Automated segmentation and characterisation of white matter hyperintensities

Carole Hélène Sudre

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Medical Physics and Biomedical Engineering

University College London

October 17, 2016

I, Carole Hélène Sudre, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Neuroimaging has enabled the observation of damage to the white matter that occurs frequently in elderly population and is depicted as hyperintensities in specific magnetic resonance images. Since the pathophysiology underlying the existence of these signal abnormalities and the association with clinical risk factors and outcome is still investigated, a robust and accurate quantification and characterisation of these observations is necessary. In this thesis, I developed a data-driven split and merge model selection framework that results in the joint modelling of normal appearing and outlier observations in a hierarchical Gaussian mixture model. The resulting model can then be used to segment white matter hyperintensities (WMH) in a post-processing step. The validity of the method in terms of robustness to data quality, acquisition protocol and preprocessing and its comparison to the state of the art is evaluated in both simulated and clinical settings. To further characterise the lesions, a subject-specific coordinate frame that divides the WM region according to the relative distance between the ventricular surface and the cortical sheet and to the lobar location is introduced. This coordinate frame is used for the comparison of lesion distributions in a population of twin pairs and for the prediction and standardisation of visual rating scales. Lastly the cross-sectional method is extended into a longitudinal framework, in which a Gaussian Mixture model built on an average image is used to constrain the representation of the individual time points. The method is validated through a purpose-build longitudinal lesion simulator and applied to the investigation of the relationship between APOE genetic status and lesion load progression.

Publication list

Published

1. C. H. Sudre, M. J. Cardoso, and S. Ourselin, "Bilayered anatomically constrained split-and-merge expectation maximisation algorithm (BiASM) for brain segmentation," in *SPIE Medical Imaging* (S. Ourselin and M. A. Styner, eds.), vol. 9034, pp. 903411–903411–7, International Society for Optics and Photonics, International Society for Optics and Photonics, 2014.

2. C. H. Sudre, M. J. Cardoso, W. Bouvy, G. J. Biessels, J. Barnes, and S. Ourselin, "Bayesian Model Selection for Pathological Data," in *MICCAI 2014* (P. G. Et al., ed.), LNCS 8673, pp. 323–330, Springer International, 2014.

3. C. Sudre, M. J. Cardoso, W. Bouvy, G. Biessels, J. Barnes, and S. Ourselin, "Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation.," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 2079–2102, apr 2015.

4. M. J. Cardoso, C. H. Sudre, M. Modat, and S. Ourselin, "Template-based multimodal joint generative model of brain data," in *International Conference on Information Processing in Medical Imaging*, pp. 17–29, Springer, 2015.

5. C. Sudre, M. J. Cardoso, and S. Ourselin, "Longitudinal segmentation of age-related white matter hyperintensities," in *The 2nd International Workshop on Bayesian and Graphical Models for Biomedical Imaging (BAMBI)*, 2015.

Under review/ In preparation

1. C.H. Sudre, B. Gomez Anson, I. Davagnanam, A. Schmitt, A. Mendelson, F. Prados, L. Smith, D. Atkinson, A. D. Hughes, N. Chaturvedi, M. J. Cardoso,

F. Barkhof, H. R. Jäger, and S. Ourselin, "Automated regional-zonal representation of white matter hyperintensity volumes on brain MRI." *Journal of Neurology Neurosurgery and Psychiatry (under review)*.

2. C.H. Sudre, M. J. Cardoso, J. Barnes, C. Frost, F. Barkhof, N. Fox, and S. Ourselin, "APOE 4 status determines WMH volume accumulation rate independent of AD diagnosis,." *Neurology (under review)*.

3. C.H. Sudre, M. J. Cardoso, and S. Ourselin, "Longitudinal segmentation of age-related white matter hyperintensities." *Medical Image Analysis (in preparation)*.

4. M. Pardini, C. H. Sudre, F. Prados, O. Yaldizli, V. Sethi, N. Muhlert, R. S. Samson, S. H. van de Pavert, M. J. Cardoso, S. Ourselin, C. A. G. Wheeler-Kingshott, D. H. Miller, and D. T. Chard, "The relationship of grey and white matter abnormalities with distance from the surface of the brain in multiple sclerosis." *Journal of Neurology Neurosurgery and Psychiatry (under review)*.

5. M. ten Kate, C. H. Sudre, A. den Braber, E. Konijnenberg, M. J. Cardoso, P. Scheltens, P. J. Boomsma, Dorret I.and Visser, S. Ourselin, and F. Barkhof, "Correlation of white matter hyperintensities in cognitively healthy elderly monozygotic twin pairs." in *VasCog conference 2016*.

Acknowledgements

I would like to deeply thank my primary supervisor, Professor Sébastien Ourselin for his guidance through my PhD, his help in putting my research in perspective, opening my horizons and for initiating fruitful collaborations. My thanks go also to my secondary supervisor Professor Nick Fox for his constructive advice and his helpful encouragements.

I am extremely grateful to my tertiary supervisor, Dr M. Jorge Cardoso for his constant support throughout these past years, his patience, availability and contagious enthusiasm. His help has truly been crucial in the evolution of my work and I benefited immensely from his advice and knowledge.

I would also like to thank all those I collaborated with and who opened my views to new perspectives and new challenges with in particular Frederik Barkhof, Josephine Barnes, Chris Frost, Beatriz Gomez Anson, Rolf Jäger, David Wallon and Ferran Prados Carrasco.

I owe a lot to my colleagues at the Translational Imaging Group and at the Dementia Research Centre whose friendliness makes these places vibrant with energy and motivation.

Finally, I would like to acknowledge the unconditional support I received from my friends and family. Nothing could have been done without their constant encouragements, motivational words, care and understanding. Special thanks to my parents, my brother and Sebastiano.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical imaging has been developed to enable the visualisation of hidden reality and investigate, describe and assess otherwise inaccessible reality of the human body. Subsequent image analysis comes naturally as a way of robustly and objectively characterising, summarising, describing and quantifying the information embedded in the obtained images. Such process is however put in place to answer clinical needs, biological questions and more globally health issues. As the title of this thesis emphasises, most of the contribution of this work can be classified at the end of the process in the attempt to robustly quantify information coming from medical images. Apart from the inherent intellectual challenge, this would however seem less meaningful without being driven by a clinical need. The rationale behind medical image analysis in this case holds behind the following questions: assuming that the observations in medical images are representative of a biological process, is it possible to provide one or different solutions to robustly quantify these observations so as to answer a clinical question? Can these objective descriptions be proven relevant and participate in a better understanding of the complex underpinning clinical reality?

This introduction tries to reflect the strong link between the three aspects of pathophysiological clinical observation, imaging and image analysis. The synergy between this three themes in the context of white matter hyperintensities (WMH) is made stronger by the fact that clinical questions and theories have only risen when neuroimaging has enabled the depiction of such abnormal observations. Besides, the denomination itself is directly derived from an imaging perspective. First, the biological situation and the needed clinical context of age-related white matter hyperintensities are detailed in Section 1.1 before addressing in Section 1.2 the importance of neuroimag-

ing and reporting the existing solutions for the visualisation of this specific condition. The challenges of quantification, particularly in terms of delineation of the investigated WMH, and the wide range of available methods (that will be more detailed in Chapter 2) are broadly mentioned in Section 1.3 before finally outlining the rationale and main components of this thesis in Section 1.4.

1.1 Brain, white matter and pathology

To describe the brain, different scales and perspectives can be adopted: cellular components, structural tissue types, anatomical regions of interest or functional connections are some of the ways the brain can be investigated. Neurons, cells that carry the electrical information across complex networks are at the core of the brain processing. They may be considered as composed of three parts: the cell body, the dendrites and the axonal extension. To interact together, most neurons use both electrical and chemical information transfer pathways. Information under the form of electrical impulses will be conveyed along the axon of the neuron until reaching its end. To "communicate" the so far electrical information, a chemical encoding is performed and the neurotransmitters released as a cellular answer to the electrical message will reach the dendrites of the second neuron via the synaptic cleft. Upon chemical reception, this second neuron may in turn proceed with the firing of a further electrical impulse. In the brain, neurons are usually well organised and the cell bodies with the dendrites packed together into what is defined as the grey matter (GM). Clinical research has so far largely focused its attention on grey matter since the cortical GM is more directly accessed and links between brain grey matter lesion and clinical outcome are naturally observed. In the popular language, grey matter as even been used as a metonymy for the whole brain with for example Agatha Christie's hero referring proudly to his "little grey cells". The reality of the brain is however much more complicated and the billions of neuronal cells could not perform appropriately without the adequate support of other brain cells called glial cells. One of the crucial aspects in the functioning of the brain lies for instance in the need for neurons to be appropriately supplied in energy via nutrients and oxygen while being simultaneously protected from any potentially dangerous compound. This complex system of protection and supply is named the blood-brain barrier (BBB).

1.1.1 Defining white matter

As defined by Malloy et al. [1], the white matter (WM), located under the cortical ribbon, represents 40 to 50% of the brain volume of a young healthy adult. It is composed by the neuronal axons and their supportive glia. Among these supportive elements, the astrocytes, cells that enact the blood brain barrier process, control the delivery of nutrients and restrict the incoming of damaging elements. In turn, another type of cells, the oligodendrocytes, are key elements to produce and maintain the myelin sheaths that coat the axons. To that aim, they develop membrane processes that encircle the axons forming multi-lammelar structures. The myelin, composed of lipids at 80%, acts as an insulator and allows for the saltatory electrical conduction along the axons, from one point of discontinuity (called node of Ranvier) to another, thereby increasing the speed of neuronal signal transmission [2]. Responsible for the connection of the different regions of the brain, white matter is therefore crucial to the transfer of information.

The myelin-coated fibres of the white matter gathered in bundles or tracts can be separated in three main subgroups according to their end-points [1]:

Projection tracts : These long tracts connect the cortex with the spinal cord, as well as the diencephalic and mesencephalic regions of the brain. The ascending tracts connect the thalamus and basal ganglia to the cortex.

Commissural tracts : These tracts serve to the lateral connection between hemispheres.

Association tracts : The cortical association tracts within an hemisphere are subclassified into two groups. The short U fibres connect adjacent gyri and are rarely affected by white matter disease. The long association tracts however are more disparate and susceptible to suffer from cerebral small vessel disease (CSVD). These tracts are likely to traverse periventricular regions and deep white matter. The longest among these association tracts are related to the frontal lobes that are a place of predilection for white matter disease.

Damages to the white matter either affecting the myelin and/or leading to axonal loss affect deeply the signal transmission and may in turn be associated with cognitive impairment. The pathophysiological explanations for such damages range from

genetic disorders (leukodystrophies) to inflammatory diseases (lupus) going through toxic or traumatic events, infectious diseases (HIV), demyelinative disorders (multiple sclerosis), or vascular pathologies [3].

1.1.2 Ageing and white matter disease

The definition of white matter disease in ageing population is difficult due to inconsistencies in the terminology [4]. In general, the term white matter disease (WMD) is used as a generic term to refer to pathological observations in the white matter with a presumed ischaemic origin as opposed to multiple sclerosis or leukodystrophies [5].

1.1.2.1 Other lesions than WMH

White matter lesions associated with cerebral small vessel disease have been classified by Wardlaw et al. [4] as small subcortical infarcts, lacunes, enlarged perivascular spaces (EPVS), cerebral microbleeds (CMB) and leukoaraiosis (LA). The latter, defined and introduced based on neuroimaging findings [6], has also been naturally named white matter hyperintensity (WMH) and will be described in a neuroimaging perspective in Section 1.2.4. Lacunes are older small infarcts that have cavitated and are fluid-filled. Enlarged perivascular spaces also known as enlarged Virchow-Robin spaces consist in the abnormal extension of the fluid space around the vessels and run along them at the interface between vessel walls and glia limitans. Normally microscopic and invisible, they can expand with age. Used for interstitial clearance towards the ventricles, and as inflammatory response pathways, their enlargement is associated with some fluid entrappement and is observed in ageing population and in multiple sclerosis patients. According to their location in the brain, pathophysiology may be different: hypertensive arteriopathy affects principally the deep (brainstem, basal ganglia and deep white matter) arterioles while cerebral amyloid angiopathy effects are mostly observed on cortical and leptomeningeal vessels [7]. Those enlarged PVS may also be a marker of BBB impairment in small vessel disease [8]. Lastly, cerebral microbleeds are believed to relate to vascular leakage of blood cells in the perivascular space.

1.1.2.2 Leukoaraiosis/WMH

Leukoaraiosis, is a term introduced in 1986 [6] to account for the non-specific changes observed in the white matter in the ageing population. Its greek roots *leuko* that refers to the colour white and here white matter and *ario* that means loose and rare, com-

bined with the action transforming suffix *-osis* was conceived to mean "rarefaction of the white matter". Reflecting the neuroimaging observations that are associated to it and will be further developed in Section 1.2.4, the term white matter hyperintensities (WMH) is indifferently used in its place. The description as white matter changes (WMC) has also been proposed.

To account for these tissue changes occurring in ageing [9], various biological explanations have been proposed such as the degradation of the myelin, the increase in extracellular water, the desaturation of the myelin lipid contributing to its instability, the damaging action of free radicals or the effects of inflammation with a reactive astrocytosis [10]. On an histological perspective, white matter changes can present a partial loss of myelin, axons and oligodendrocytes, a slight increase in the number of other glial cells, a stenosis of the arterioles as well as the presence of macrophages [11]. In the ageing process, the myelin degradation has been associated with myelin pallor, the loss of myelinated fibres and the deformation of myelin sheaths [9].

Historically, WMH have commonly been considered as a consequence of the partial ischaemia of the tissue. Unable to obtain their required survival components, neurons and oligodendrocytes progressively decay leading to the degradation of the myelin sheath [12]. In general, damage to the blood vessels through wall thickening, lumen narrowing and vessel stiffness may impair the tissue perfusion therefore preventing its metabolic needs from being answered in a timely and appropriate manner. Higher biological markers of hypoxia (lack of oxygen) have been reported in WMH [13] and arteriolosclerosis has been observed as the most important pathological finding [14].

Increasing evidence shows the relationship between an impaired vascular endothelium and the occurrence of WMH [15] and markers of increased BBB permeability, that not only control the tissue supplies but also protects the brain against toxic attacks have been reported [3]. Indeed, depending on the location, impairment of the endothelium can lead to an accumulation of plasma protein in the vessel walls, damaging the structure of the vessels at the smooth muscle cells level or by the deposition of fibrin. These alterations may directly affect blood flow auto-regulation with subsequent inappropriate tissue perfusion especially when evolving at later stage to vessel stiffness and lumen narrowing. Alternatively, if leakage of plasmatic compounds in the perivascular space occurs, for example at the level of the capillaries for which no smooth muscle layer

exists, the toxicity of the leaked compound through vasogenic edema is potentially threatening to the surrounding tissue. Additionnaly, if red blood cells are extravasated from the vessels, alterations in the iron metabolism can induce oxidative stress thereby promoting the damage [12]. Implications of the iron metabolism are further backed up by the recent discovery of the independent association of the gene for hemochromatosis with severe white matter changes [16]. Further impairment at the venule level especially around the ventricles has been observed with venous collagenesis [13].

The blood-brain barrier is known to be more permeable with increasing age which is in line with the fact that age is a major risk factor for the occurrence of WMH and that markers of BBB dysfunction have been observed in WMH. Even in the absence of risk factors, the potential for vasodilation reduces with age and the expression of inflammatory compounds increases. However, the exact pathways leading to BBB disruption in normal ageing are still unclear [17]. In the case of WMH, feed-forward loops of increasing damage may occur, explaining the exponential accrual of observed lesions [15]. For instance, hypoxia is known to contribute to a higher BBB permeability that, as mentioned before can lead to further vascular damage and inadequate tissue perfusion. Further inflammatory mechanisms as a response to the presence of toxic proteins or in response to ischaemia have also been observed to disrupt the structure of tight junctions at the BBB level [17].

By choice, the terms leukoaraiosis and WMH are indefinite enough to encompass the large variety in appearance, spatial distribution and potential histopathological explanations. With respect to location, varied hypothetical pathways have actually been suggested but obtaining robust and reproducible characterisations of such location properties is in itself a challenging task. Lesions are classically separated between those that are located continuously to the ventricular surface (periventricular WMH or PVWMH) and those that are not (deep WMH or DWMH).

DWMH are generally associated with the ischaemic explanation for leukoaraiosis [18] while the PVWMH can be substratified according to their appearance, with different pathophysiological explanations. In the milder cases of periventricular (PV) caps around the frontal horns evolving to a thin pencil line around the lateral ventricles lining toward smooth halos [16], the observed areas of hyperintensities appear to be due to some discontinuity in the ependymal lining delimiting the ventricles or to a

loosening of the fine fibre tracts. These changes contribute to the increase in extracellular space and thus to the water content of the region [11, 18, 19] not affecting directly the signal transmission through myelin deterioration. Therefore, in periventricular regions, the observed hyperintensities are not discriminative for myelin degradation or increased water content due to BBB disruption and should not be completely associated with myelin pathology [20]. The leakage of cerebrospinal fluid (CSF) in the brain parenchyma causing this increase in water content may not be corrected if the blood-brain-barrier is altered and the reabsorption process hindered. Irregular PVWMH are in turn mostly explained by a chronic hemodynamic insufficiency and a fibrohyalonic process.

Characterising the lesion spatial distribution is made more difficult by the fact that the damage to the white matter evolves in time. This may for instance lead to the coalescence of periventricular damage and deep white matter lesions. As underlined by Yoshita et al. [21], the problem is thus not only volumetric: a fourth dimension, time, must also be included. In the study of pathologies, the question of longitudinal evolution is crucial to clarify potential causal relationships.

In the evolution of white matter disease, a partially consistent pattern has been described. Hyperintensities seem to first appear at the horns of the lateral ventricles, then all around them before occurring in the deep white matter and the basal ganglia [1]. Maillard et al. [22] showed the trend of the lesions to extend from existing damage in vulnerable neighbouring regions. Moreover, the rate of appearance of new white matter signal change is associated with the initial volume of lesions such that more severe cases will develop more rapidly [23, 24].

1.1.3 White matter changes, risk factor and clinical correlates

On a clinical perspective, in addition to age, some risk factors, such as diabetes mellitus, hypertension or smoking, have been associated to the presence of white matter changes (WMC) [1, 15, 25, 26]. Although thought to affect mostly the arterial part of the vasculature, with emphasis on the deleterious effect of hypertension and other vascular risk factors such as diabetes mellitus and smoking, WMH have also been observed in populations free of vascular risk factors [15]. Attempting to map locally WMH and risk factors, Rostrup et al. [27] demonstrated a clearer association of hypertension in the

DWM regions while age was strongly associated to the most common areas of WMH that is the periventricular horns and lateral bands.

Although at first considered as clinically silent and benign, evidence is now pointing to a relationship between WMH and deleterious clinical outcomes. With respect to cerebrovascular damage, WMH have been shown to increase the risk of stroke in the general population [28]. In terms of cognitive abilities, the preexistence of WMH appeared to negatively affect cognitive results one year after a stroke [29]. Furthermore, WMD can be linked to cognitive decline especially with respect to processing speed disorders [30]. This observation strengthens the hypothesis that the clinically relevant WMH are those in which myelin deterioration and in turn tract disruption is involved [20, 31]. Most of the complaints were related to the executive functions with decreased processing speed, erroneous goal formation, planning or organisation [16], motor disturbances with gait disorders, problems in postural control or urinary continence [16, 32]. Due to the high concentration of white matter fibres in the frontal regions related to these functions, the probability of these regions to be affected in white matter degradation could indeed be higher and therefore explain this observation [10]. To a lesser extent, some memory deficits, and recognition difficulties have also been reported [16] as well as mood disorders such as depression [18].

Using the above mentioned distinction between PVWMH and DWMH, PVWMH have been more easily associated to executive function and processing speed deficiencies while DWMH have been mostly highlighted in relationship to mood disorders [18, 26, 33–35]. The association of mood disorders to DWMH is explained by potential disruptions to the tracts in the frontostriatal circuits, known to be related to mood regulation [11]. Such a distinction is still controversial [36] and the reviewing studies about the location of the WMH show a high heterogeneity in methods, including the definition of the WMH, the choice of the used neuroimaging technique, the method of WMH quantification and of neurocognitive assessment [35]. When looking at the link to the risk of developing dementia, the role of WMH location remains unclear [37]. However, in most of the studies, the total WMH burden has been associated with a higher risk of developing Alzheimer's disease (AD) and other dementia types for the general population but has not been found significant when predicting the transition from Mild Cognitive Impairment (MCI) to AD [37]. Associated with cortical

GM atrophy in the normal ageing population [38], two complementary hypotheses can be advanced with respect to the occurrence of white matter changes [18]. On one hand, the destruction of grey matter (GM) cells could lead, by Wallerian degeneration, to the decay of the associated fibres. On the other hand, disruptions at the axon level could propagate towards the GM by a retrograde deafferentiation process.

Studying more specifically the possible relationship between AD and the presence of white matter damage, various hypotheses have been presented to explain a potential link between the expression of both pathologies. It appears that in AD cases, white matter abnormalities are more prominent in posterior than anterior regions [1], compared to normal controls, thus inferring that a different mechanism is induced by AD in posterior regions. However, Radanovic et al. [25] conclude that WMH occur independently from the AD pathology but that the risk of cognitive decline in association with WMH is increased when amyloid deposition is observed, and they consider the presence of WMH as a sign of increased vulnerability to cognitive decline. This conclusion can in some ways be linked with the controversial hypothesis of the existence of a threshold above which silent white matter damages become symptomatic [1, 37]. Ertens-Lyons et al. [14] found an association between a higher WMH load and AD and three explanations were proposed for this finding: first the Wallerian degeneration induced by the loss of cortical grey matter driven by the AD process was causing white matter disruption; second an ischaemic cause to axonal injury may lead to tangle formation and eventually to neuronal degeneration; third a common mechanism could lead to both, ischaemia causing both white matter damage and neurodegeneration. Additionally, as amyloid increases the permeability of the BBB, white matter changes linked to BBB disruption might be further explained in AD pathology [15]. In the case of cerebral amyloid angiopathy (CAA), when amyloid plaques are deposited on the vessel walls and often observed in AD, WMH are more prominent [39].

Longitudinally, a larger increase in lesion burden is further associated with a higher brain volume loss [40], while smoking appears to accelerate the progression [41]. In terms of links with cognitive outcome, a larger WMH increase seems to be related to the absence of remission in depression [42].

All these hypotheses, theories and clinical associations with respect to WMH rely however on the observation and quantification of the lesion burden. Such relationships

would have been impossible to draw without the development of relevant neuroimaging solutions.

1.2 Medical imaging and WMH

1.2.1 Medical imaging and pathology

In order to understand, evaluate and ascertain pathological processes happening in the body, different strategies have been applied. The most invasive one, consists in putting outside what is inside. Drawing samples from tissues or fluid (blood or CSF) is possible *in vivo* but provides only a partial and limited knowledge about the reality. Moreover, these useful snapshots of a biological situation do not inform on the relative organisation of tissues. In the special case of biopsies, the randomness of the draw, may actually even mislead the diagnosis. Even in most extreme cases when the analysis is performed post-mortem, at which stage a dissection of the relevant tissues is possible, the changes in physiology may modify the structures and the applied mechanical constrains further affect the tissue organisation. Although in some cases, partial information on a pathological situation can be acquired without the sense of sight, using the sense of touch to assess swollen ganglia, abnormal tumour growth or broken bones for instance, the brain, encased in the solid skull remains inaccessible.

Medical imaging, in making visible what naturally would remain hidden to sight has allowed new and crucial insight in understanding the organisational structure and functioning of this organ as well as the potential abnormal situations that can be encountered. Abnormalities as a whole in terms of imaging, can refer to structural presentation such as shrinkage of the hippocampus or enlarged ventricles, but also unexpected intensity signals as observed in tumour masses, necrosis or WML. With respect to abnormalities, the natural assumption is that these observations are linked to a pathological process affecting the functioning of the organ and contributing to detrimental clinical outcomes. However, it must be underlined that drawing causality relationships is extremely difficult and most of the conclusions are merely associations and correlations that in time prove useful and relevant clinically.

1.2.2 Imaging the brain

For the brain, medical imaging in its different modalities relies on the differential properties of the present tissues that either respond differently to a stimulus (X-ray, magnetic field) or behave differently with respect to a source of stimulus (Positron emission tomography radiotracers). Magnetic Resonance Imaging (MRI) has been proven as a modality of choice to image differences between the soft tissues that compose the brain. In brief, it consists in analysing how tissues answer in time to different types of magnetic excitations. During an MR acquisition, a main magnetic field is applied continuously and the spins of the protons naturally align to it. Upon excitation by transient magnetic pulse oriented in a perpendicular plane, spins will align to this new field and rotate in phase at a given resonance frequency. Once the excitation pulse has subsided, spins will naturally tend to go back to their original state, governed by two main tissue properties. The T1 relaxation property also known as spin-lattice relaxation reflects the level of energy released by interactions between the protons and the surrounding molecules. The T1 relaxation time is then the time required for the protons to reach back 63% of their initial energy. In turn, the T2 relaxation property also known as spin-spin relaxation expresses how fast the excited spins dephase due to magnetic interactions between protons. In particular it helps evaluating the density of macromolecules in a tissue and the coherence of the proton organisation. With respect to the T2 property, additional inhomogeneities in the main field such as those due to susceptibility-related field distortions may strengthen the magnetic interactions. Such distortions are in particular due to the presence of blood or iron in a tissue. The association of both causes of magnetic interactions is summarised as the T2* characteristic. According to the way the magnetic pulse sequences are designed, different contrasts can be achieved exposing the varied and complementary tissue properties. With T1-weighted images (also denoted T1w or T1), thought to provide good structural contrast between tissues, the T1 relaxation properties are favoured compared to the T2 characteristics, thus emphasising the different levels of energy interactions between protons and surrounding molecules. Highly organised with large molecules, fat tissues present a short T1 relaxation time. In T1-weighted images, these tissues will then appear with a higher signal than tissue in which energetic interaction with surrounding molecules is lower, as in the CSF, that appears dark. The high proportion of lipid in the com-

position of the white matter therefore contributes to its higher signal on T1-weighted (T1) images, aligned with its autopsy derived original denomination. In the case of T2-weighted (T2w or T2) pulse sequences, the emphasis is put on the property of magnetic interactions between protons while in PD-weighted images also referred to as intermediate weighted image, the proportions of protons in the tissues are imaged. In turn T2* weighted and susceptibility weighted (SW) images convey also the information of the internal sources of magnetic distortion. More recently, Fluid Attenuated Inversion Recovery (FLAIR) pulse sequences have been developed to provide T2-weighted contrast while nulling the signal of free water.

1.2.3 Imaging white matter disease

White matter lesions (WML), globally defined as any type of signal abnormalities observed in the white matter appear in a wide range of pathologies as mentioned in Section 1.1. Due to the high variability of their appearances, their systematic study is extremely challenging, all the more so that neuroimaging findings often do not allow for the discrimination between pathological explanations [3, 9, 26]. To differentiate multiple sclerosis (MS) and age-related WMH, due to the similarities in appearance, Fazekas et al. [43] developed new criteria based on location pattern, size and evolution to distinguish between the two pathologies.

Ischaemia related injuries on the white matter can present themselves along various patterns among which the most common are the leukoaraiosis and the lacunes [1,4]. Standardised descriptions and the respective image acquisition modalities of choice have been described by Wardlaw et al. [4] to help discriminating between ischaemic neuroimaging findings related to cerebral small vessel disease (CSVD).

Small subcortical infarcts The modality of choice for the detection of recent small subcortical infarcts is diffusion weighted imaging (DWI). With a diameter inferior to 20 mm, they appear afterwards with a decreased intensity on T1 and increased in T2 and FLAIR in the region of a perforating arteriole.

Lacunes They are round or ovoid in shape, usually fluid filled, thus with a signal similar to the CSF and their diameter range from 3 to 15 mm. On FLAIR images, lacunes usually appear as an hypointense centre surrounded by a rim of hyperintensity. When it is not completely fluid filled, such lesion can appear totally

hyperintense on a FLAIR image. Usually, the old lesions appear smaller than the recent ones because of an *ex vacuo* effect in the old lesions and a tendency to swelling in the new ones.

Enlarged perivascular spaces When enlarged, they resemble lacunes in much of their appearance since they are also seen as fluid-filled cavities with CSF like signal on all sequences. They differ however in their shape since they usually follow the course of a vessel whereas lacunes appear ovoid in shape and are usually larger than 3 mm in diameter [4]. They are mostly prominent in the lower basal ganglia.

Cerebral microbleed A cerebral microbleed appears generally hypointense on T2* image with a round or oval shape in small areas with a maximum diameter of 10 mm. Their dark appearance is due to the accumulation of hemosiderin-laden macrophages at these locations.

1.2.4 Imaging white matter hyperintensities

1.2.4.1 Impact on structural images signal

White matter hyperintensities, as their denomination indicates, are observed as bright on certain MRI pulse sequences. These abnormalities have actually been described when noticed in medical images and were underlined as potential focus of research thanks to the advances in neuroimaging. This terminology may however be slightly misleading since other lesion types (*e.g* subcortical infarcts) may also be reported with hyperintensities. Historically such bright spots seen on brain images of T2-weighted or PD-weighted images were denoted as UBO (undefined bright objects). In age-related cases of white matter damage, deterioration of the myelin sheath, increase of water in the extracellular volume near the ventricles contribute both to the change in the ratio between fat and water in the tissue and so between densely packed molecules and free protons. In periventricular regions, the inability of structural images to distinguish between those origins has been substantiated by histological studies suggesting that the hyperintense signal observed do not only correspond to the myelin deterioration [20,24] but also to a higher water content. The change in water/fat ratio [1] in turn directly affects the T1 and T2 relaxation time and the tissue properties appear to move towards the properties of the CSF. This is the reason why, in cases of T2-weighted contrast, such

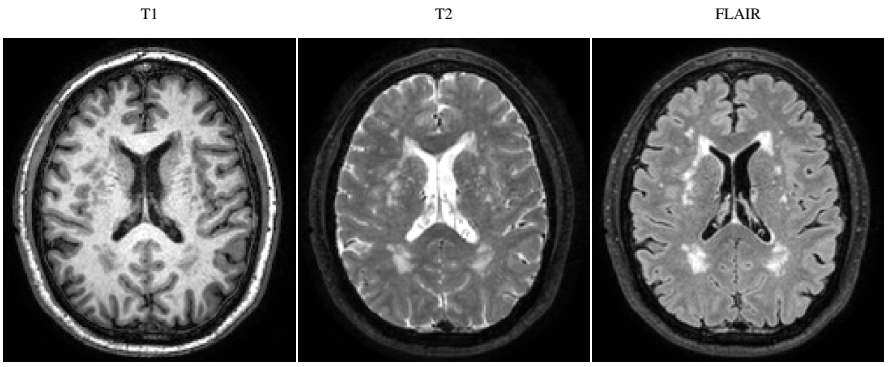


Figure 1.1: Example of leukoaraiosis on T1, T2 and FLAIR MR images.

damage is observed as brighter than the healthy tissue. However, if the water proportion increases in such lesion, it is usually not completely fluid filled which justifies why it is still observed as bright on FLAIR images except in most severe cases. T2-weighted, PD-weighted and FLAIR sequences, all sensitive to the damage to the white matter have therefore been particularly used for the visualisation and assessment of WMH. Figure 1.1 illustrates the presence of leukoaraiosis on T1, T2 and FLAIR images.

Trying to study WMH in a normal elderly population in order to avoid finding dementia confounders to WM deterioration, Murray et al. observed that a high FLAIR signal was related to a higher vacuolation linked both to myelin loss and decrease in small vessel density [44]. The vacuolation enables indeed the accumulation of interstitial fluid that contributes to increase the FLAIR and T2 signal.

As mentioned in Section 1.1.1, it appears that the pathology often starts from the ventricular lining. Separating between CSF and lesions in T2 and PD-weighted images can therefore become quite challenging since both present a high signal. The main strength of the FLAIR sequence relies in its ability to null the CSF signal while preserving a bright signal for the hyperintensities. In contrast, these regions of hyperintense signal in T2 and FLAIR usually present iso to slightly hypointense signal in T1 depending on the severity of the damage and even though used by Leritz et al. [45], T1-weighted images are usually not well suited for the quantification of leukoaraiosis [46]. Very hypo-intense regions in T1 would correspond to regions of complete degradation of the tissue as with a complete infarct. In the definition of leukoaraoiosis itself, some have even considered that the corresponding T1 intensities should be isointense to the healthy white matter [47]. Nonetheless, with the evolution of the disease and the increasing damage to the tissue the distinction may be sometimes difficult to draw.

1.2.4.2 Focus on FLAIR sequence

FLAIR sequences, although becoming the structural modality of choice for the assessment of WMH thanks to its ability to clearly differentiate lesions from ventricles, suffers from various drawbacks. The rationale behind the use of medical images is, as mentioned earlier for the general field of medical imaging, its assumed reflection of an existing damage to the brain tissues. Links between histopathological findings and imaging observations on FLAIR pulse sequences have however demonstrated that FLAIR images tend generally to overestimate the extent of lesions [48–50]. Besides, the intensity observed on FLAIR scans cannot be directly related to the severity level of the lesion [51,52] contrarily to T1 or T2 acquisition sequences. Moreover, it must be underlined that the imaging hyperintense signals are non discriminative with respect to the underlying pathological process. Therefore, in different diseases, the presentation of lesions may be strongly similar as can be the case for MS lesions and age-related WMH.

When relying on FLAIR images to observe WMH, one must be aware of possible normal hyperintense confounders and artefacts that may affect their depiction. Regarding artefacts, those are more prominent at 3T than at 1.5T due to increased field inhomogeneities and magnetohydrodynamics [53]. They include pulsatile CSF flow within the ventricular system or at subarachnoidal spaces, magnetic susceptibility at the floor of the frontal lobe due to air and bone structure, spatial misregistration of the anterior cerebral artery or problem in the nulling of the CSF due to the presence of metallic foreign objects. An increased incidence of pulsatile flow artefacts has been observed for expanded ventricular system [54], thus being more problematic when studying an ageing population potentially presenting enlarged ventricles. The effect increases at places with fast pulsatile flow such as the Sylvan aqueduct or the fourth ventricle. Besides, those artefacts are not limited to ventricular system and may overlay adjacent structures. The fine delineation of lesion may be further hindered by shine through effects, present as increased signal intensity at the border between parenchyma and ventricles or between cortical ribbon and external CSF. All these artefacts observed in 2D acquired FLAIR images are luckily mostly resolved by the use of 3D FLAIR images [53, 55] but such acquisitions are still not widely performed in clinical settings.

At the time of the development of FLAIR pulse sequences, studies have moreover

reported the potential lower sensitivity to lesions in the infratentorial region of the brain while other tissues than damaged white matter may present similar MR properties and result in similar intensity signal. Additionally, normal findings in FLAIR images such as a bright septum pellucidum [56,57] may further hinder the definition of the WMH to assess and encourage to an increased caution in its definition [23,46,58]. As of yet, very few directives have however been given on how to delineate and distinguish normal and abnormal hyperintense findings appropriately [59]. This question of border delineation is further thwarted by the continuity observed in the damage severity between normal appearing and hyperintense tissue. This continuity is further highlighted by the observation of a progressive degradation of the normal appearing tissue neighbouring the lesions [22].

1.2.4.3 Alternative choices of pulse sequences

Other MR pulse sequences initially developed for other applications than the study of age-related white matter changes may bring further information relative to the characteristic of white matter damage. Magnetic transfer imaging (MTI), that measures the ratio between mobile and bound water is for instance highly sensitive to demyelination and used for multiple sclerosis (MS) to evaluate damage severity. Diffusion tensor imaging (DTI), that provides measures of the ability of the water to diffuse freely in the tissue can also be used to assess how water is bounded through anisotropy evaluation and consequently reflect white matter integrity [24]. It suffers however from image distortion, high sensitivity to movement, low resolution and the difficulty to register it to high resolution structural images [1]. Its true link to white matter integrity remains controversial [9] but it has been used to study the normal appearing white matter (NAWM), shown to be more vulnerable at the neighbourhood of existing lesions [23, 60]. A combination of modalities even only structural may bring a more accurate information on what to consider as WMH [49, 61] and might further help distinguishing lesion subtypes.

1.3 WMH and quantitative image analysis

1.3.1 Need for quantitative assessment - visual grading and volume

Whatever the clinical hypotheses tested or the associations that are being drawn between white matter damage in any type of pathological presentation, a quantitative assessment must be used to ascertain the extent of the damage. Two main options are available for that purpose: visual grading scales and volumetric measurements. In the case of age-related WMH, grading scales have been designed to provide a quantification of the severity of the damage over the whole brain. Although those visual gradings are fast to obtain, they suffer from grade non linearity, their lack of sensitivity to small changes and their susceptibility to ceiling and flooring effects [18,62]. In fact, they convert continuous data into categorical elements. Moreover, the training to apply them properly is strenuous [1] and inter- and intrarater variability that varies according to lesion load [63] adds to the uncertainty. The heterogeneity of the designed scales in terms of location separation, grading methods, lesion characterisation or modality on which they are applied, has been put forward as the reason for discrepancies in clinical association results [64]. Others find however that some of those scales correlate well with each other and the correspondences allow for a partial translation between scales [65]. These opposite findings regarding the influence of the choice of visual scales can be explained by the improvement in image acquisition that make the lesion burden easier to assess. Nonetheless, it has been reported that the finer grading scales, defined on a more local level appear to be more relevant when looking at the correlations between WMC and cognition [62, 66].

Generally those scales correlate well but not linearly with the volumetric measurements of the white matter damage [1, 62]. The Fazekas scale has even been shown to give almost equivalent estimates of the WMH burden [58]. The main discrepancies in the correlations are either due to the misdetection of subtle hyperintensities as lesion or the disproportionate importance acknowledged in grading scales by the presence of small subcortical lesions compared to the induced relative lesion volume change. For longitudinal analysis however, volumetric measurements appear to be more reliable than classical visual scales [67]. Scales designed for the purpose of longitudinal assessment were shown to perform better when compared to cross-sectional scales [66].

Furthermore volumetric measurements of the white matter hyperintensities appear to enable the assessment of more subtle cognition changes than visual scales [62].

If those observations support the use of volume measurement compared to visual scales, the non-linear correlation with volume measurement could also be related to the hypothesis that additional lesion descriptions could further refine the information brought by the overall volume lesion load [61]. Linking different types of lesion characterisation with the rationale of the used visual scales would further help to better define the most important features of the lesions and thus make their analysis more clinically meaningful.

1.3.2 Importance for lesion segmentation - need for automation

Quantitative measurements based on the delineation of tissues such as volumetric estimations need to be robust and accurate to ascertain subtle and valid clinical associations. As far as lesions are concerned, the importance of their accurate segmentation is doubled. Not only can the information coming from their delineation be invaluable in terms of clinical association, but this delineation can impact greatly the accuracy of other quantitative measurements based on tissue delineation. It has been reported for different pathologies (MS, AD), that not accounting for the existence of lesion could affect healthy tissue volume measurement such as grey or white matter [68–70]. Moreover, obviating the presence of lesions may also have a deleterious effect on other image analysis processes such as image registration [69]. Manual segmentation of lesions, a time consuming and cumbersome task, suffers from inter and intrarater variability [71, 72] and is therefore hard and costly to implement in large clinical studies. Even though decreasing the time required [71], semi-automated methods of WMH segmentation remain highly dependent on the operator with still a strong time constraint. These techniques may involve observer [73] or automatic [74] thresholding, contour evolution [75] or texture analysis [76] from an initial rough drawing and may potentially be followed by a manual editing refinement stage [77]. With the advances in MR imaging, 3D images are now more commonly available to assess lesions. Reported to more accurately depict white matter damage such images contribute however to an increase in time expense for operator-based lesion segmentation methods. Thus, and even more so in the perspective of large population studies, the development of reli-

able, robust and reproducible automated white matter lesion segmentation techniques has become crucial.

1.3.3 Automated WMH segmentation: goals and strategies

The arduous task of WMH segmentation reflects the need to obtain as reliably and objectively as possible a satisfying delineation of WMH. Ideally, the solution would allow to find all lesions (sensitivity) but no more (specificity), while being consistent in the definition of their borders across varied conditions of resolution, noise and artefacts.

As underlined in Section 1.2.4, WMH are areas of bright signal present in the white matter. This description in itself provides strong constraints on the task at hand in terms of signal and spatial criteria. Focusing for a moment on the word hyperintensity, the prefix hyper meaning over, beyond or in excess, refers to a comparison and therefore to the potential inclusion of a certain degree of subjectivity in the definition of the described lesion. Among others, the question of the definition of the level of outlierness arises along with the reference to the normality. These issues need to be addressed somehow consistently when trying to automate them. Among the reported challenges to detect WMH, the fact they can occur everywhere with a high variability in shape and location is not the least. Furthermore, the intensity signature of such lesions can be confused with healthy grey matter especially at the border of the lesions thereby leading to misclassifications. These observations of mixed intensities are due to the presence in a single voxel of both normal and damaged tissue. This problem, related to the image resolution, is called partial volume effect (PVE).

As will be detailed in Chapter 2, numerous techniques have been developed for the automated segmentation of white matter hyperintensities in various pathological situations, mostly MS and age-related WMH. Their grouping is made challenging by the variability in the number of steps considered and in the relative importance of detection and potential refinement stages described in each method. Indeed, if one considers with great care the question of an appropriate detection that would require only (if any) minor corrections [78–82], others develop with more emphasis means of refining the segmentation to avoid false positives ([83,84]) or consider both these stages with equal importance [85–88]. Such distinction makes blurrier the border between the core of the method and what could/should be considered as postprocessing. The classical grouping

between supervised and unsupervised methods is also blurred and does not correspond to the same definition according to reviews. In the review by Lladó et al. [50] for instance, any method including knowledge from examples of a population is classified as a supervised method. Therefore, methods using statistical atlases, that are maps of tissue probabilities obtained from an average of population segmentations, are considered as supervised. The degree of supervision could also be subject to discussion since, even if not using explicitly a database of manual segmentations, knowledge and comparison with existing examples may influence methods designs and parameters choices. In complement, data-based tuning of methods parameters [89, 90] could be argued to belong to supervised methods.

A possible, although artificial, line of demarcation between methods can be drawn between discriminative methods that derive a criteria to distinguish lesions from other tissues and those developing generative models of the data accounting for lesions. Methods are classified as belonging to the generative group if they attempt to model how all observations are distributed even if these involve a secondary non-generative step. In the first group, the discriminant factor may be a single intensity threshold, or derived from the study of more complex feature vectors with respect to a manual lesion segmentation database and appropriate classifiers. Such methods differ in the type of features and classifiers. In turn, generative methods that are robust to the presence of lesions describe strategies that model the outlierness of given observations with respect to healthy tissues. They may either adopt non-parametric strategies that consider the full data sample or model the data parametrically through classes described by a limited number of characteristics. Many options have been developed to that respect, either modelling lesion classes explicitly or excluding outliers from the representation of healthy observations.

Apart from the definition of intensity abnormality level, inclusion of anatomical knowledge either within the chosen method or as a postprocessing step is essential to limit the amount of false positive detection. A popular way of including such information both in discriminative [85, 91–95] and generative [78, 79, 96–100] methods is through the use of healthy tissue atlases obtained from the averaging of tissue segmentations of an healthy population and aligned to the cases of interest. With respect to postprocessing and final refinement of lesion segmentation, some strategies do take

pride into the fact that all risks for false positives are accounted for [78, 101]. Such a step may affect considerably the segmentation results.

However, WMH rarely stand alone and are often combined to other pathological observations. For instance, the reported clinical importance of measures of atrophy in both ageing and MS and the observed impact of lesions in such measurements highlights the need for a joint robust segmentation of healthy tissues [102]. In both ageing and MS, the picture of the pathology is not complete if only taking into account WMH since other lesion subtypes also contribute towards the pathological description. Thus, apart from the technical differences between the existing strategies, other aspects in the statement of the problem, such as the need for high-throughput in the processing, the type (binary or fuzzy) of segmentation, the need for complementary outputs (segmentation of healthy tissues, of other lesion subtypes) or generalisation potential impact the initial methodological choices.

Such choices relate in particular to the availability and need for specific pulse sequences. At one end of the spectrum, monospectral strategies generally consider the FLAIR image as their modality of choice. Indeed, compared to the T2w and PDw images on which WMH also appear bright, the main advantage of the FLAIR sequence is the nulling of the CSF signal which enables for a good separation between lesions and CSF, important to delineate periventricular lesions. The promoters of the monospectral strategies [103–107] argue that it allows for results free of registration error since no alignment between images is required. Additionally, it is less costly in terms of scanning and usually computationally less demanding. However, FLAIR images suffer from some drawbacks such as the presence of flow artefacts, a tendency to overestimate the lesions and a low sensitivity in the infratentorial regions as mentioned in Section 1.2.4.2. With the increase in computational power weakening the processing speed argument, multispectral signal is nonetheless often considered as more capable of correctly determining the lesion extent and is often used to avoid or correct for false positives.

1.4 Motivation of the work

1.4.1 Rationale

Rarely is a methodological automated solution developed to answer a single condition scenario problem. In itself, the word automated connotes an idea of large applicability and strong generalisation potential. More intricate and subtle clinical questions requiring larger study populations warrant such methods in the case of WMH quantification. In any automatic technique, learning, tuning and rules are biased by the knowledge introduced to answer a specific question. This issue is especially sensitive for discriminative classifiers relying on databases of specific segmentation examples but exists also for methods that incorporate heuristic rules. Strategies that decouple the robust modelling of the data from the expected output have therefore the advantage of being transferable. Although discriminative methods based on segmentation examples may enable a perfect reproduction of manual delineations, intrinsic human errors may simultaneously be learned and reproduced. With respect to generative techniques, those that model lesions explicitly or integrate their description in their data explanation may overlook the existence of other types of outliers or lesions while those that consider unexpected observations as a whole may overconstrain their data representation and miss finer descriptions. Decoupling the complete data modelling from the application process may thus overcome these potential limitations. Neither the complete (inlier+outlier) data joint modelling nor the subsequent application tuned solution should however be considered as an end-result. Considering finer representations, parallel applications in the perspective of the clinical needs should indeed further allow for a deepening of the understanding of the underlying pathologies.

1.4.2 Thesis Overview

This thesis presents a generic model selection framework, named BaMoS (Bayesian Model Selection), which models multimodal images with abnormal intensities using a hierarchical Gaussian mixture model (GMM). The end-result of this model selection process is further applied to automatically segment and characterise white matter hyperintensities. Chapter 3 will introduce the theoretical background of GMM and their optimisation in the Expectation-Maximisation (EM) algorithm. Previously developed improvements to the GMM are then described, namely the correction for intensity inho-

mogeneities, the introduction and relaxation of *a priori* anatomical knowledge through statistical atlases and the addition of neighbourhood context constraints via a Markov Random Field. To these commonly applied variations, a further constraint on the covariance matrices is developed. This chapter ends with a short review on model selection in GMM with a focus on split and merge strategies. Chapter 4 first explains the development process that led to the model selection framework in Section 4.1 before detailing the final hierarchical model (Section 4.2) and explaining how the variations detailed in Section 3.2 can be adapted to this framework (Section 4.2.2). The implementation details are then presented in Section 4.3. Chapter 5 focuses on applying the framework to the study of white matter hyperintensities. How the final GMM obtained through BaMoS can be used to provide the white matter lesion segmentation is detailed in Section 5.1. Before validating BaMoS within the context of lesion segmentation, the assessment measures based on a gold standard reference are described and discussed in Section 5.2. An internal validation is first carried out in Section 5.4 before comparing the developed algorithm to other available lesion segmentation methods in Section 5.5 using both simulated and clinical data. As noted in Section 1.3.1, lesion volumes may not gather the full extent of the information and specific biological questions may require finer descriptions (cf Section 1.1.2.2) related in particular to the location and spatial distribution. Chapter 6 is thus devoted to the development and applications of a systematic patient specific location scheme enabling the description local lesion characteristics, then applied in different contexts and at both population and individual levels. If the location information can bring further insight into the understanding of WMH pathology at a specific time, longitudinal changes are crucial to the establishment of causality relationships especially if associated to specific lesion patterns.In that perspective, Chapter 7 is dedicated to the longitudinal extension of BaMoS then validated on a purpose-built lesion simulator (Section 7.3) and on clinical data (Section 7.4). Finally, Chapter 8 attempts to summarise the work presented in this thesis and presents avenues of further investigation.

Chapter 2

State of the art in WMH segmentation

As evoked in Section 1.2.4, different pathologies can result in similar abnormal intensities on MR images. This is the case for the unexpected hyperintense white matter signal observed both in age-related leukoaraiosis and in multiple sclerosis (MS). Compared to MS cases, healthy tissues in ageing are less contrasted and WMH present smoother borders. These differences have been highlighted as the reason why a technique developed for one pathology is difficult to transfer to the other [49]. Some methods have however been successfully validated in both situations [89, 100] or marginally adapted as with the method developed by Schmidt et al. [90] and revalidated by Maldjan et al. [108]. Reviews on WMH segmentation imaged through MRI have been recently published focusing either on the MS problem [50, 109, 110] or on the elderly case [111]. Therefore, methods developed both in the context of an elderly population (*e.g.* [49, 81]) or MS subjects (*e.g.* [78, 79, 88, 112]) will be mentioned in this chapter.

Across the many different algorithms developed to automatically segment white matter hyperintensities, one important commonality lies in the need to preprocess the data before analysis. Section 2.1 is therefore devoted to this aspect of the segmentation pipelines. The other end of the process, that tackles the refinement of preselected potential lesions and the correction for misclassifications, is expanded in Section 2.4.

Following the demarcation line described in Section 1.3.3, methods are divided between purely discriminative and partially generative strategies. Among the discriminative techniques that attempt to draw a line between what should or not be considered as WMH, methods are classified according to the choice of features used to obtain the lesion discrimination. The decision criterion can be based on a single intensity threshold [94, 103, 113] possibly combined with image enhancement techniques [95, 114],

include spatial [80, 82, 91] or texture based [115] information or be built from the comparison between patches [112, 116].

Since this work develops a generative model of data in presence of outliers with application to WMH segmentation, emphasis is given in this review to methods that model the image globally trying to account for the presence of lesions, or more generally of outliers. Lesions can either be explicitly (Sections 2.3.1 and 2.3.2.1) or implicitly (Section 2.3.2.2) modelled using class centroids [78] or established probabilistic distributions such as Gaussians [86] to parametrise the data. Alternatively partially non-parametric solutions can be developed to model the data in presence of abnormal observations [117].

Lastly, for all the techniques developed in the perspective of an application to WMH segmentation, the question of validation and evaluation with respect to criteria of generalisability, robustness, specificity, sensitivity and clinical relevance and their comparison to existing solutions is crucial. Guidelines for validation have been detailed by García-Lorenzo et al. [110] and the different strategies encountered are briefly evoked in Section 2.5.

2.1 Preprocessing

Preprocessing of the data is an inherent part to the lesion segmentation process. Such a step may include skull-stripping, correction for intensity inhomogeneity, images alignment or intensity normalisation [92, 112, 118]. In some cases, the segmentation of other tissues using available brain segmentation tools such as SPM or Freesurfer is also considered as a preprocessing step [119–121]. Although the impact of changes in the parameters of the preprocessing steps is rarely assessed, Zijdenbos et al. [122] underlined its strong standardisation as a factor of major importance for the generalisability of proposed techniques. Since normalisation is needed when intensity features are compared across images, Steenwijk et al. [118] compared different normalisation strategies and showed that this choice impacted both the optimal parameter and the measures of overlap with manual segmentations in the case of their k-nearest neighbour procedure. Skull-stripping and fat removal are usually performed to improve the intensity normalisation, limit the computational cost [80] and reduce the risk of including tissues whose intensity distributions are close to the investigated WMH [123].

This is especially important for FLAIR images [124] in which artefacts around the eyes may appear [125]. Although some strategies use in-house methods for skull-stripping [99, 104, 117, 117, 126, 127] or base it on thresholding applying the Otsu technique that minimise the within class variance between background and brain tissues [114, 124], available algorithms such as BET [128], SPECTRE [129] or MBRASE [130] are also commonly applied. Despite its influence on normalisation, no analysis is available to evaluate the influence of these skull-stripping strategies on the subsequent lesion segmentation.

Another source of problems, often taken care of in a preprocessing step, is the presence of smooth variations in intensity across the image. This artefact, known as intensity inhomogeneity (IIH), bias field (BF) or gain field [131] is mostly due to imaging instrumentation and the nonuniformity of the applied magnetic fields. The most accepted way of representing image intensity affected by such an effect in the presence of noise is $y = \alpha x + \xi$ where ξ refers to the noise, x to the true intensity and α to the bias field effect. Correcting for this artefact has been shown by Johnston et al. [132] to improve the lesion segmentation. When included in generative models, linear combinations of basis functions (trigonometric or polynomial) are often used to model the smooth variations [78, 79, 133]. Otherwise, a very large proportion of the methods described in this review, tend to use the N3 method developed by Sled et al. [134] in the preprocessing of their data. It consists in deconvolving the histogram and finding an intensity shift to which a smooth function is fitted to minimize the signal entropy.

Additionally, enhancement and denoising of the images may be applied in strategies that usually do not include a parametric modelling of the data at hand. Enhancement techniques may however lead to an increase in the level of noise as noticed by Boudraa et al. [135] when using histogram equalisation techniques. Furthermore, when the contrast varies across the image, local solutions should be used. One of the most common enhancement techniques is the application of an anisotropic diffusion filter that has the property of smoothing homogeneous regions while enhancing the edges. Such a filter has initially been introduced by Perona and Malik [136] for 2D images and its 3D implementation is available in the ITK library, an open source C++ library dedicated to medical imaging research (`www://http.itk.org`), which contributes to its popularity. Its iterative effect on the image I, with λ a user-defined parameter is

expressed as

$$\frac{\partial I}{\partial t} = \operatorname{div}\left(g\left(\| \nabla I \|\right) \nabla I\right)$$
$$\text{with } g(x) = \frac{1}{1+\frac{x^2}{\lambda^2}}$$
$$\text{or } g(x) = \exp\left(-\frac{x^2}{2\lambda^2}\right)$$

If the use of such enhancement procedure is quite common in the case of MS pathology, Dyrby et al. [81] have underlined its possible limitations in the case of age-related lesions for which lesion edges are smoother.

Alignment between images, known as registration, can also be required as a pre-processing step in three main situations. First an intra-subject alignment or coregistration of the images is often required when multiple pulse sequences are used for a single subject. Indeed, even if acquired during the same scan session, the subject may have moved, or the resolution be different between pulse sequences. The reference space chosen varies across methods; this choice may depend on the further processing but the resolution may also play a role. Additionally, the choice of reference space for the registration has to be taken into account when further evaluating the accuracy of the methods. Second, a general knowledge built over a population in a specific space, often referred to as template space, has sometimes to be transferred to the individual case. Such generic knowledge may for instance take the form of tissue statistical atlases, that are maps giving an a priori probability for each voxel to belong to a certain tissue. An alignment transformation between template and individual space is then needed to propagate the information. Lastly, some methods require for all subjects in a population to be spatially comparable resulting in an inter-subject alignment into a common space. Very commonly, the MNI (Montreal Neurological Institute) population template [137] is chosen as space of reference [112, 118, 138–140].

In all situations, registration implies the resampling of interpolated intensities for the modified images. Apart from the occasional registration error leading to inconsistent intensity pairing, the resampling may in particular blur tissue edges. In the registration problem, one tries to estimate the transformation that enables the best match between the images. Depending on the allowed degrees of freedom and type of mod-

ifications, two main families can be distinguished. On one hand, the linear registrations can be expressed as a matrix multiplication transformation applied globally to the image with a maximum of 12 degrees of freedom in tridimensional cases: rotation, translation, scaling and shearing in each of the directions. On the other hand, non-rigid or elastic transformations allow for local image modifications and can be associated to many more degrees of freedom. When dealing with images bearing lesions, linear registration is commonly favoured over non rigid solutions due to the effect of lesions on registration as evoked by Guizard et al. [112] and reported by Skika et al. [141]. The non-linear solution applied by Tomas-Fernandez et al. is claimed not to be affected by the presence of pathology [87].

2.2 Discriminative methods

2.2.1 Unidimensional feature: intensity thresholding

Based on the denomination of such lesions, the most natural solution to separate WMH from healthy tissues appears to be the determination of a threshold above which all voxels of a FLAIR image should be evaluated as potential candidates. In the method of Jack et al. [103], an histogram is drawn for each slice of the skull-stripped, bias field corrected, anisotropic filtered FLAIR image, and the characteristics of the central peak of the distribution are derived. Based on a dataset of manual segmentations, these characteristics are used in a linear regression to determine the appropriate threshold for WMH. A database of manual segmentations is also used by Gaonkar et al. [125] to obtain a global optimal threshold this time and correction for false positives is based on a probabilistic map. Ong et al. [94] also consider a threshold on the full FLAIR histogram but do not use a database to derive the optimal threshold to apply. The lesions are selected from a 1.5 standard deviation from the third quartile of the assumed normal tissue distribution and corrected for false positives with a WM mask derived from the T1 based tissue segmentation. It is also the search for an optimal threshold that warranted the work by Yoo et al. [105]. Using the parenchymal histogram from the result of the SPM8 segmentation, and modelling the variation of the z-score threshold with respect to the volume of WMH, a database of manual segmentations is used to estimate the parameters. False positive correction is realised based on a WM mask template registered to the image.

Using only the GM histogram of the FLAIR image obtained after segmentation of the T1 image, works from Roura et al. [120] and de Boer et al. [113] bear a lot of similarities. After thresholding, only lesion clusters with a high enough proportion of neighbours segmented as healthy WM are retained. In both cases the various threshold parameters (histogram and neighbourhood proportion) are tuned based on a set of manual segmentations. The most important difference between both methods lies in the way the initial tissue segmentation is obtained; Roura et al. use the outcome of the SPM5 segmentation while De Boer et al. implement an atlas-based k-nearest neighbours (kNN) stategy.

2.2.2 Image transformation and region growing

The previously described methods focus on how to obtain an optimal threshold for the WMH discrimination on the FLAIR histogram. Others have dedicated their efforts to the image enhancement in order to increase the contrast and allow for an easier WMH discrimination.

Samaille et al. [95] process the FLAIR image through serial steps of anisotropic diffusion filtering and watershed segmentation in order to delineate homogeneous intensity regions. Lesion regions are then selected based on a thresholding of the initial FLAIR histogram while the result of the T1 segmentation is used to correct for false positives.

In the work of Zhong et al. [104] a conservative threshold is first applied to the FLAIR WM histogram. In order to remove elements from the GM, a high pass filter is applied to the image. The remaining lesion seeds are finally used to initialise a region growing algorithm.

In a similar perspective, Pattino-Correa et al. [114] also select seeds to initialise a region growing framework. Such seeds are obtained from the image, smoothed with a mean filter and enhanced with gamma correction and removal of the background mean.

2.2.3 Multi-dimensional features and classifiers

Classification techniques use features extracted from a set of labelled examples to train a tool (the classifier) that will then decide under which label a newly presented data point should be categorised. Techniques vary widely both in the extracted features and in the training and classification method. In order for comparable features to be

extracted from the images, normalisation steps are necessary both for intensity and spatial information. Normalisation of the spatial information is for instance realised by aligning all images to a common space.

Sweeney et al. [142] concluded from their comparison between the impact of features and classifier choice that the feature design impacted more the segmentation result than the choice of classifier. Popular classification techniques in WMH segmentation range mainly among k-nearest neighbours (kNN) [80, 92, 118, 125]), support vector machine (SVM) [82, 140], random forests [91, 115] and neural networks [81, 143]. In the kNN technique, a distance metric is chosen and the k samples of the training set that are the closest to the sample to classify are selected; the average of their labels is attributed to the sample of interest. With the SVM, it is the hyperplane separating with the highest margin positive and negative findings that is looked for. The label of an incoming sample is obtained according to the side of the hyperplane on which it will lie. In the case of the random forest, decision trees are built for which at each node a binary feature based decision is defined so as to maximise a given criterion representing the consistency of the data in each subclass. Building many such trees based on slightly different training samples and/or feature sets leads to a random forest. For a given new sample, the label decision is then the average over all results given by the decision trees. Finally, with neural networks, the association between features and labelling is the result of an activation function on the combination of functions of the weighted features. Such a combination is learnt so as to minimise an objective cost function.

As for the feature vectors, they can be based only on intensity [92], or encompass also voxel spatial location [80], texture characteristics [115] or derive further information from statistical priors [91, 118]. With the recent increase in computational power, patch-based methods are lately gaining in popularity.

2.2.3.1 Voxel-wise features

Zijdenbos et al. [143] propose an artificial neural network to segment healthy tissues and WMH based on T1, PD and T2 images but since the classifier is trained from samples selected by an operator for each image, this can only be considered as a semi-automated method. In the k-nearest neighbours algorithm derived by Anbeek et al. [80], all modalities are registered to the FLAIR image and all tissue intensities are

gathered into the feature vector. In order to account for the spatial location of the lesions, coordinates of the studied voxel are added as features. Improving on this initial method, Steenwijk et al. [118] incorporated in the feature vector statistical probabilities of tissue label and normalised the spatial information to the MNI space. In parallel, Wu et al. [92] also considered a kNN algorithm to separate various tissues and lesion subtypes instead of giving a binary answer. The output of the kNN was then corrected through the application of a template derived segmentation. Introducing anatomical spatial information through atlases appeared to consistently improve the segmentation as mentioned in the comparative study led by Kamber et al. [93]. Dyrby et al. [81] studied the impact of feature vectors used in a neural network classifier developed for a multi centre study. The comparison between feature vectors underlined the relevance of complementary intensity information from modalities other than FLAIR especially in the case of low lesion load.

2.2.3.2 Explicit context enriched features

In a random forest framework, Geremia et al. [91] use as a basis the feature vector proposed by Steenwijk. et al. [118] transformed by mean filtering to account for noise. Context features are included as the difference between the local feature value and the mean over a large distant patch or the mean over a region symmetrical with respect to the mid-sagittal plane. In the solution of Cabezas et al. [85], the features considered include, in addition to the intensities and *a priori* atlas tissue probabilities described by Steenwijk et al. [118], the probabilistic values of an outlier map. They complement the vector with meta-features similar to those proposed by Geremia et al. [91] obtained as the difference of the feature value and the mean of other features over large distant regions. Sweeney et al. [139] obtain context enriched features thanks to the smoothing of the images at multiple window scales for all modalities. With a different weight given to each feature, the logistic regression used for classification allow for a relative weighting of the modalities. Using coregistered T1 and FLAIR images, Ithapu et al. [115] include explicit texture features from a patch surrounding each voxel of interest and investigated the performance of support vector machine (SVM) and random forest (RF) classifiers on such feature vectors.

2.2.4 Patch-based approaches

With the increase in computational power, patch-based techniques applied to lesion detection have recently been more widely investigated. In those techniques, the information attributed to one voxel comes from a patch over its neighbourhood.

After image smoothing via a Gaussian kernel, Lao et al [82] use a SVM on feature vectors that do not contain any direct spatial information but in which all the intensities in the neighbourhood of a given voxel are concatenated. A similar feature vector is used by Damangir et al [140] where, in order to avoid the imbalance between sample sizes in the healthy and lesion class, an active sampling is applied in a series of cascading SVM that aim at pruning the non lesion samples from the sample pool. In the approach of Jog et al. [138], a regression tree is learnt over the patch intensities and meta-features close to those described by Cabezas et al. [85]. A lesion membership is calculated as the average of the labels per voxel of each patch in the resulting leaf node, that is then thresholded and corrected for false positives with a WM mask derived from the T1 segmentation.

Similarity measures between patches can also be used in the classification process. Introduced initially for denoising purposes, the non local mean (NLM) includes such similarity evaluation. With the voxels j belonging to the neighbourhood Ω of voxel i, the NLM $\hat{x}(i)$ is expressed as

$$\hat{x}(i) = \frac{\sum_{j\in\Omega} w_{ij} x(j)}{\sum_{j\in\Omega} w_{ij}} \text{ with } w_{ij} = \exp\left(\frac{\| P(x(i)) - P(x(j)) \|_2^2}{h^2}\right)$$

where w_{ij} is a measure of the similarity between the patches centred in i and j and h is a smoothing parameter.

After alignment to a common space, Guizard et al. [112] first select in their database the most similar subjects to the case of interest. For each voxel i, a weighted label is obtained from a variant of the NLM applied using patches of neighbouring voxels coming from all selected subjects. Optimal results are realised when combining T2 and FLAIR modalities and including mirrored images to the database. In turn, Mechrez et al. [116] use for each studied patch the NLM weighting to the selection via kNN of the k most similar patches of their database. Spatial information is further introduced by the estimation of the labelling distance between patches.

In the work of Deshpande et al. [144], a sparse dictionary learning technique is applied on the patches to learn the characteristic of a healthy and a lesion class and each tested patch is then classified globally as healthy or abnormal. Majority voting over all overlapping patches at one voxel gives the final lesion classification. The dictionary learning patch-based procedure developed by Roy et al. [145] performs the learning using a different set of patches selected with a kd-tree for each new presented patch. The lesion membership of the patch central voxel is the weighted sum of the labels of training patches.

2.3 Generative models

Generative models for lesion segmentation vary greatly in their presented strategies and different classification criteria can be highlighted to differentiate them. A great source of variability lies for instance in the way pulse sequences are considered in a pipeline ranging from the monospectral approaches [124] to the fully multispectral ones [79] going through those that incorporate the information progressively [135]. In the methods described here, the robust generative model that takes outliers into account is often not the only step in the pipeline resulting in the lesion segmentation. When present, additional steps may rely on thresholding [101], gradient based elastic evolution [86], fuzzy connectivity [117] or graph-cuts [87].

As the denomination suggests, parametric methods are based on the estimation of parameters needed to explain the distribution of the observed data while non-parametric solutions consider the full-sample distribution. For lesion segmentation, some parametric methods may attempt to explicitly model both normal and abnormal observations while others may keep the modelling of outliers implicit. Alternatively, non-parametric models may be used to ascertain the presence of unexpected intensities. Parametric methods gather models based for instance on Fuzzy C-Means (FCM) clustering or Gaussian Mixture Models (GMM).

2.3.1 Explicit modelling with fuzzy-C Means clustering

In the fuzzy-C Means clustering, classes are parametrised based on their intensity centroids. The soft segmentation is based on the minimisation of an energy function ac-

counting for the distance between sample and centroids features with the form

$$E = \sum_{n=1}^{N} \sum_{j=1}^{J} w_{nj}^{m} \parallel y_n - c_j \parallel^2,$$

where w_{nj}^{m} is the membership of voxel n to class j, c_j the centroid feature (usually intensity) of class j and y_n the intensity at voxel n. The fuzziness parameter m describes how fuzzy the segmentation will be, with the limit of crisp delineation when m tends towards 1. The weight expression is obtained as

$$w_{nj}^{m} = \frac{1}{\sum_{k=1}^{J} \left(\frac{\parallel y_n - c_j \parallel}{\parallel y_n - c_k \parallel} \right)^{\frac{2}{m-1}}}.$$

Boudraa et al. [135] apply a 3 class FCM on PD images enhanced by histogram equalisation and power law and use the corresponding T2 intensities to select potential lesion and CSF classes. Reiterated on the selected elements, a second FCM provides a more accurate selection that is then corrected for false positives based on minimum size and adjacency to brain border. Admiraal-Behloul et al. [49], use FCM in parallel on T2, PD and FLAIR images with different numbers of clustering classes and the resultant information is then combined through a fuzzy inference system based on intensity linguistic rules. Template prior atlases, used to initialise the centroids of the FCM, are also further employed to improve the false positive correction.

One of the difficulties in fuzzy-clustering is the incorporation of spatial consistency in order to avoid misclassification errors due to the presence of noise in the data. In the strategy of Anitha et al. [146], information on the spatial variability is included to the FCM framework and shown to improve the lesion segmentation applied on an elderly population for which FLAIR images contrast is previously enhanced. The FCM energy functions has also been adapted in order for instance to model the bias field as in the method of Gao et al. [133]. Additionally, in this work, spatial consistency is introduced in the energy function through the application of a non-local mean regulariser over the distance to the corrected centroids. In the method developed by Shiee et al. [78], the energy function includes the bias field modelling and both statistical and topological atlases are incorporated to ensure spatial consistency. Local modality

weighting, minimum size rules and distance constraints are added to the model so as to account for potential false positive detection.

2.3.2 Lesion segmentation with Gaussian mixture models

Since the Rician noise observed in MR is classically modelled as close to Gaussian, Gaussian mixture models (GMM) are a very common type of parametric models. They consist in modelling the data as a weighted sum of Gaussian distributions, each of them representative of a different tissue type. The distribution of the data $\mathbf{Y}$ is thus expressed as

$$f(\mathbf{Y}) = \sum_{k=1}^{K} \pi_k \mathscr{G}(\mu_k, \Lambda_k),$$

where K is the number of Gaussian components in the model, μ_k and Λ_k the parameters of the normal probability distribution $\mathscr{G}$ of class k and π_k the corresponding mixture weight. To obtain a labelling of the image based on such a model, a common solution consists of introducing hidden labels and applying the expectation-maximisation (EM) algorithm described by Dempster et al. [147]. The EM alternates between the estimation of the posterior responsibilities of the labels given the current parameters (expectation step) and the update of the Gaussian parameters (maximisation step) in order to increase the log-likelihood of the complete data (intensities and labels). The theoretical background of the EM algorithm is detailed later on in this thesis (Section 3.1).

The EM is however known to be sensitive to the presence of outliers [148]. Solutions either modelling explicitly the lesion distribution or robustly estimating the healthy tissue parameters have thus to be adopted.

To establish the level of abnormality of a given observation with respect to a tissue class, the Mahalanobis distance is often chosen [101]. In a multispectral view, the Mahalanobis distance for a sample $\mathbf{y}_n$ to the mean of a Gaussian distribution of parameters $(\boldsymbol{\mu}, \Lambda)$ is expressed as

$$d_{\text{Mahal}}(\mathbf{y}_n, \boldsymbol{\mu}, \Lambda) = \sqrt{(\mathbf{y}_n - \boldsymbol{\mu})\,\Lambda^{-1}\,(\mathbf{y}_n - \boldsymbol{\mu})^T}.$$

In order to ensure spatial consistency in the labelling and limit the effects of noise, as explained by Zhang et al. [149] and further detailed in Section 3.2.3, Markov Ran-

dom Fields (MRF) can be introduced to limit labelling variability by incorporating information from the neighbourhood classification in the probabilistic estimation of the labels [79, 89, 107].

2.3.2.1 Explicit modelling with Gaussian mixture models

After skull-stripping and bias field correction, Simões et al. [89] model a FLAIR image using a 3 component Gaussian mixture model with CSF, GM+WM and WMH classes initialised based on the histogram analysis. The EM algorithm is associated to a neighbourhood filter in order to enforce regularity in the segmentation and the resultant WMH map is thresholded before false positive correction. The two extrinsic model parameters (size of the neighbouring filter and probabilistic threshold for the binary segmentation) are jointly tuned based on a dataset of segmented lesions. Khayati et al [107], propose a solution based on the combination of an adaptive mixture model (AMM) with an iterative Bayesian classifier that includes a Markov Random Field for neighbourhood consistency. Incorporating sequentially the observations to the model, a new Gaussian component is added each time the observed sample cannot be properly modelled by the existing distributions. The Gaussian components are then grouped under three umbrella classes whose parameters are iteratively optimised using the AMM and an MRF to constrain the probabilities. A generalised extreme value distribution (EVT) models the lesions in combination with a 2-component GMM for the healthy tissues in the work of Wang et al. [123]. In this case, the hard segmentation between EVT and GMM is iteratively used to progressively discard samples in the image and recompute the healthy tissue parameters until convergence. The adaptation to a multispectral case simply implies the use of a 3-component GMM.

Exploiting the different pulse sequences sequentially, Schmidt et al. [90] use the segmentation performed on the T1-weighted image to create beliefs maps that are then thresholded and used as seeds to initialise the lesion segmentation. The FLAIR image is modelled with a 3-component GMM for the healthy tissues combined to a gamma distribution for the lesion class. To ensure spatial consistency, only voxels neighbouring lesion labels can be reclassified from healthy to damaged tissue in the optimisation process. Wango et al. [121] also use the pulse sequences sequentially obtaining first a T1 weighted segmentation. Similarly to Simões et al. [89], the FLAIR image is

first modelled as a 3 component GMM and a discriminative threshold for WMH is determined. After correction for false positives, the T1 segmentation is used to select the WM +DGM volume and a 2 component GMM is optimised on the FLAIR image after being initialised with the parameters of the first model. This step not only allows to constrain the lesions to the WM but make the estimation of the lesion class more robust due to its higher proportion in the data to model. Indeed, the argument of instability due to the small available number of samples is classically underlined as a potential caveat of methods modelling lesions as a distinct separate class with a defined distribution function.

Among the multispectral techniques, Wei et al. [98], building on the method developed by Warfield et al. [150], consider T2 and PD sequences acquired in a dual-echo sequence and apply an EM with 4 classes initialised with the statistical model of a single subject. The initial GMM is postprocessed via morphological operations and application of a template deformation performed with gradient analysis so as to consider only lesion within the white matter. In the method developed by Forbes et al. [100], that includes the use of an MRF and atlas tissue priors, a local informative weight is attributed to each used modality (T1 T2 or PD or FLAIR). From the 3 class model obtained after convergence of a variational Bayes EM, candidate lesion regions are defined and used to initialise a 4-class model optimised in a similar way. The possibility of the existence of other types of outliers in the data is however not considered.

Using more than 3 Gaussian components to model the healthy tissues has been also proposed in the context of lesion segmentation with multispectral data. In the solution designed by Freifeld et al. [86], many spatial Gaussian components are fitted and classified into intensity tissue clusters. Spatially 1/20 of the voxels are used as seeds for the spatially localised Gaussian distributions and the intensity parameters are initialised based on a K-means algorithm. Each of the components contains individual spatial information and relates to a global tissue intensity characteristic. The parameters are optimised through an EM algorithm for all the Gaussian components. Starting with three intensity clusters, heuristic intensity rules are used to reclassify after convergence lesion components under a fourth intensity cluster. After this correction procedure, the EM is processed again and the lesion segmentation is further improved by the use of a curve evolution framework on the locally convex lesion regions. In con-

trast, Harmouche et al. [151] challenge this idea of intensity stability across the brain and consider that separate GMM including lesion Gaussian components should be optimised per homogeneous region of interest. In this case, the regions are predefined and atlas priors are used to ascertain a good initialisation. Spatial consistency is enforced via the application of an MRF. Instead of predefined regions, Galimzianova et al. [152] developed a method defining the homogeneous regions and applied existing lesion segmentation algorithms to the specified regions to obtain a stratified model. Such regional separation can be put in parallel to methods that process the infratentorial region or the deep grey matter separately [99, 126].

2.3.2.2 Implicit modelling in Gaussian mixture models

If some strategies do choose an explicit model distribution for the lesion intensities, others may not impose such constraint. In order to avoid biasing the segmentation of healthy tissues, robust solutions that limit the influence of the outliers on the estimation of the healthy tissue parameters must then be applied [148].

For instance, the trimmed-likelihood estimator (TLE) only uses the fraction 1-h (h user-defined threshold, known as trimming parameter) of the data best represented by the model to estimate the parameters [96, 97, 101, 124, 153]. Aït-Ali et al. [97] use healthy subject atlases to initialise the parameters of a 3-component GMM. After convergence of the TLE, a Mahalanobis distance threshold is applied to select candidate lesions, that are further refined based on intensity heuristic rules. In this case, the use of multimodal data appeared to improve the final segmentation result. Bricq et al. [96] augmented this solution with the introduction of context information through the use of hidden Markov chains and statistical atlases. Wang et al. [124], determine the optimal value for the trimming parameter h based on a database of manually segmented cases and corrected the selected candidate lesion voxels by morphological operations.

Instead of considering all and only voxels excluded from the estimation of healthy tissue parameters as potential lesion voxel candidates, the robust parameter estimation and segmentation results may be used to characterise a secondary threshold for the lesion segmentation [101, 127]. In order to avoid using atlases of healthy subjects to initialise the parameters, García-Lorenzo et al. [101] implement a hierarchical initialisation based on the modes of the histograms that are combined with the TLE. The

lesions are then segmented based on a Mahalanobis distance threshold and heuristic rules of intensity, size and neighbouring tissues. An adaptive trimming obtained from the Mahalanobis distance value is used by Dugas-Phocion et al. [127] to estimate robustly the healthy parameters. The lesion segmentation on the FLAIR image is obtained by applying a Mahalanobis distance threshold based on the parameters estimated for the normal tissues. Refined with the addition of partial volume classes [154], the scheme has further been used by Souplet et al. [126] with a postprocessing pipeline that allows after a generous selection of potential lesion locations for a refinement based on a WM mask. In those last methods however, both outlier and some partial volume classes may prove relevant to the final lesion segmentation. In order to limit the influence of statistical atlases in regions where errors are likely to occur, Cabezas et al. [153] introduce a similarity map summarising potential registration errors. Similarly to Bricq et al. [96], a probabilistic threshold is used instead of the Mahalanobis distance to adapt the fraction of trimmed voxels. The resulting robust GM FLAIR histogram was then thresholded before final refinement. Compared to the thresholding methods relying on preliminary segmentation described in Section 2.2.1 [95, 113, 120], the methods mentioned in this section insist on how to make their model more robust to the presence of outliers rather than on how to select the relevant threshold. In a different perspective, García-Lorenzo et al [88] used the output of the applied TLE combined with heuristic intensity rules to initialise a graph-cut separating normal appearing tissues from lesions so as to provide seeds of source and sink.

Rather than trimming samples from the likelihood estimation, Van Leemput et al. [79] introduce a typicality weight to downweight the influence of outliers samples towards the parameter estimation. The MRF used to promote spatial consistency across the image is adapted in order to promote the association of lesion voxels with WM and vessel voxels with CSF. Outlier samples that do not satisfy the heuristic intensity criteria for being vessels or lesions are not downweighted. Segmentations of healthy subjects are used in the approach proposed by Tomas-Fernandez et al. [87] to locally inform if the segmented tissues are healthy or not and contribute to the downweighting of the outliers in the model parameters estimation. After selecting the elements that contribute the least to the model, a graph-cut algorithm is used to separate MS lesions from other potential related source of outlierness such as cortical lesions, iron

deposition in the basal ganglia or T1 hypointense lesions. Compared to the work from García-Lorenzo et al. [88] for which the graph-cut procedure is applied on the full data, it is here only executed on the selected lesion candidates.

2.3.3 Mixed parametric and non-parametric methods

Sajja et al. [117] propose a method combining parametric and non parametric methods in order to segment not only lesions but also healthy tissues in the brain. At a first stage, after skull stripping, bias field correction and contrast enhancement through anisotropic diffusion filtering, 2-dimensional Parzen window non parametric technique is used on T2 and FLAIR images in order to separate lesions, CSF and parenchyma. From a set of N sample points of a given class, the density Parzen-Window estimate for the observation x is expressed as

$$P(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h_N^d}K\left(\frac{x-x_n}{h_N}\right)$$

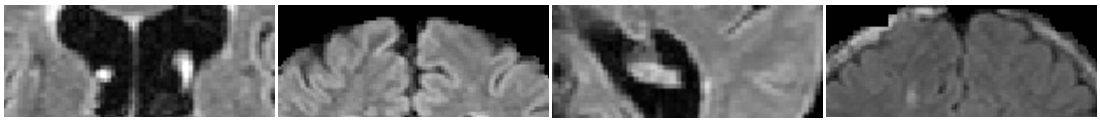
where K is a density kernel, h_N the window bandwidth and d the number of dimensions used. The label attributed is the one corresponding to the highest obtained density estimate. A threshold on the ratio between PD and T2 intensities is used in order to remove potential false positives within the brain and too small clusters are excluded. Using the T2 and PD images, GM and WM are segmented with an EM algorithm incorporating spatial context information thanks to an MRF. Islands of GM within the WM are then reclassified as lesion. The method is however considered semi-automated as it depends strongly on the sample points chosen by an expert at the initial stage of the Parzen window process. Closely to this work, Datta et al. [99] have also used a Parzen classifier to separate CSF, lesions, and parenchyma based on T2 and FLAIR images. Intensities are normalised across images in order to avoid having to choose for each case a new feature set for the Parzen window training. The result of the non parametric process is then used to initialise two EM algorithms including an MRF on T1 and T2 images to separate GM and WM. One of the two EM is devoted to the separate analysis of the cerebellum. Fuzzy connectivity rules are finally applied to refine the lesion delineation. García-Lorenzo et al. [155], use the mean shift clustering to create regions of homogeneous intensities. Once obtained, these regions are combined based on the difference in mean intensity between neighbouring regions. Regions depicting

lesions are finally selected according to an intensity threshold derived from the results of a TLE applied to the multispectral data and their adjacency to WM.

Table 2.1 summarises the characteristics of the previously described generative lesion segmentation strategies.

2.4 Postprocessing

As underlined in Section 1.2.4.2, in most cases, WMH coexist in the image with other sources of hyperintensities that can therefore be mistaken for lesions. The CSF flow artefacts at the border between GM and CSF or within the ventricles, remaining skin, fat or bony structure, or elements of the choroid plexus are among the probable suspects. Figure 2.1 gives examples of such hyperintensities confounders. Even though spatial features or anatomical knowledge may be incorporated in the presented strategies, many methods do require such a step to provide a relevant segmentation. The detection of those elements is actually so important that for some methods the line between the core principle of the method and the postprocessing aspect becomes unclear while others focus more on the ways of correcting for false labelling than on the detection itself [83, 84]. Theoretically, such a step is quite debatable: if it is needed, does not it mean that the features used or the model developed are suboptimal? Practically, consistent problems arise across datasets and common solutions are proposed. Due to noise, isolated voxels may for instance be segmented as lesion. To correct for this effect and select only biologically meaningful elements, a threshold over the minimum size of the lesions is often applied either volumetrically [101, 107] or slice by slice [135]. Steenwijk et al. [118] explore this aspect of the lesion refinement in its impact on the segmentation overlap but conclude that it leads to very small changes in the evaluation measures. To correct for the segmentation as lesions of flow artefacts and non brain tissues at the outer region of the brain, a common strategy is to eliminate all detected



a b c d

Figure 2.1: Examples of hyperintensities cofounders: a) Ventricular flow artefact; b) Outer brain flow artefact (bright cortical ribbon) c) Choroid plexus d) Remaining outer tissue after skull stripping.

		Application	Modalities	Preprocessing	Lesion segmentation	Anatomical	Spatial consistency	Refinement	Ref
Non Parametric	Boudraa et al. [135]	MS	T2 PD	Histogram equalization	2 FCM on PD with T2 masking	No	No	No	NA
	Anitha et al [146]	Elderly	FLAIR	Bright contrast stretching	FCM with geostatistical inclusion	No	Spatial variability	No	NA
	Gao [133]	MS	T1 T2 FLAIR	BF Skull - Reg	4 classes Energy Min	No	NLM	No	T1
	Shiee [78]	MS	T1 FLAIR	Skull - Reg	4 classes Energy Min Topological clustering Distance rules	Atlases	Topological atlas	No	NA
	Admiraal-Behloul [49]	Elderly	T2 PD FLAIR	Skull - Reg	3 for T2 3 for FLAIR Fuzzy Rules	Atlases	No	Size	PD
	Khademi [106]	MZ	FLAIR		Edge based Partial volume	No	No	Atlas WM No	
Parametric	Ait-Ali [97]	MS	T2 PD T2 PD	Skull - BF - Reg	TLE	Atlases	No	No	
	Bricq [96]	MS	T1 T2 PD T1 FLAIR	Skull - BF - Reg	TLE adaptive with MRF	Atlases	HMRF	Size Atlas WM	
	García-Lorenzo [101]	MS	T1T2PD T1T2FLAIR	Skull - BF - Reg	TLE Detection (proba hyper)	No	No	Size Neighbourhood	T2
	García-Lorenzo [88]	MS	T1T2PD	Skull - BF - Reg	NABT and Graph cut over everything	No	No	WM mask Brain border	
	Wang [124]	MS	FLAIR	BF - Skull	TLE with learnt choice of h	No	No	Size Morph CSF	
	Dugas-Phocion [127]	MS	T1T2PDFLAIR	Skull - BF - Reg	NABT Adaptive TLE	Atlases	No	out/in Neigh Atlas WM	
	Souplet [126]	MS	T1T2FLAIR	Skull - BF - Reg	Gaussian PV classes + Outlier Robust Threshold from pure GM	Atlases	No	WM seg	
	Cabezas [153]	MS	T1 T2 PD T1 T2 FLAIR	Skull - BF - Reg - ADF	TLE / Threshold	Atlases	No	Size Neighbourhood WM seg	
	Van Leemput [79]	MS	T1 T2 PD	Skull - Reg	Outlier class / downweighted EM MRF	Atlases	MRF	No	
	TomasFernandez [87]	MS	T1T2FLAIR	Reg - BF - Skull	Downweighting from healthy models Threshold on fit and graph cut on selected candidates	Healthy models	No	No	
	Simoes [89]	MS / Elderly	FLAIR	Skull - BF	3 class GMM. Threshold over probabilities	No	Context filter	Morph CSF Seg	
	Khayati [107]	MS	FLAIR	Normalisation	AMM and MRF	No	MRF	Size	
	Wang [123]	Elderly	FLAIR T1T2FLAIR	Reg - Skull - BF	2 or 3 GMM with EVT	No	No	No	
	Schmidt [90]	MS / Elderly	T1FLAIR	BF - Reg - Skull	T1 segmentation Beliefs map. Probabilistic GMM lesion growing		MRF	No	T1
	Wango [121]	Elderly	T1T2FLAIR	Reg - Skull - BF	Double GMM 1 overall with T1 seg init1 only from WMDGM		MRF in preprocessing	Atlas WM Size	
	Wei [98]	MS	T2PD	Smoothing Skull	4 classes	MS based patient	PVEC	Morph Atlas	
	Forbes [100]	MS / Elderly	T1T2PD T1T2FLAIR		2 steps VEM local weighted modalities	Atlases	MRF	Deformation No	
	Freifeld [86]	MS	T1T2PD T1T2FLAIR	Normalisation	Multiple local Gaussians gathered per tissue intensity 2 steps (3 classes Reclassification 4 classes)	No	Spatial Gaussian	Curve evolution	
	Harmouche [151]	MS	T1 T2 FLAIR	Reg - Brain BF - Intensity Normalisation	1 GMM per region (lesion as Gaussians)	Atlases	MRF	No	
Mixed	Sajja [117]	MS	T2 PD FLAIR	Skull - ADF BF	LesionCSFPar (Parzen) GMWM HMRF EM	Samples from operator	MRF + Fuzzy connectivity	Border Ratio int WM Seg	T2
	Datta [99]	MS	T1T2 FLAIR	Skull - BF - Reg - ADF - Intensity normalisation	LesionCSF Parzen window . GMWM EM HMRF. Separation Cerebellum and DGM	Atlases + Samples	MRF	Border Atlas CGM Morph Atlas Ventr	T2
	García-Lorenzo [155]	MS	T1T2PD T1T2FLAIR	Skull - BF - Reg - Denoising	TLE NABT Mean Shift Region selection	No	Mean Shift	Size Neighbourhood	

Table 2.1: Summary table of generative lesion segmentation techniques

voxels within a certain distance to the mask border. Alternatively, morphological operations of erosion and dilation on the lesion maps are sometimes sufficient to get rid of artefacts at the border of the cortical ribbon or within the ventricles [89,94,99,119,121]. Lastly, one of the stronger constraint is that they should belong to the WM and previous segmentation of healthy tissues can contribute to the correction for false positives. Proportions of neighbouring healthy tissues have for instance been proposed to ensure the appropriate context of lesions [113, 120, 153]. Since healthy tissue segmentation can be affected by the presence of lesions, this solution may however become a chicken and egg problem. In fact, as mentioned in Section 1.2.4, WMH signal may sometimes be confused with GM signal. Therefore, lesions may be segmented as islands of GM within the WM. Morphological operations on WM masks such as hole filling have been popularly used to avoid these false negative misclassifications [95, 98, 126].

2.5 Methodological validation

Once a method has been implemented, it needs to be validated according to criteria of sensitivity, specificity, robustness to noise and artefacts, reproducibility and clinical relevance.

2.5.1 Ground truth vs Gold standard

In the specific case of lesion segmentation, different aspects contribute to make the evaluation of the results controversial and difficult. Validation usually requires the availability of a ground truth against which to compare the obtained segmentation. Such a ground truth is however a fleeting concept that is rarely available. More commonly, a gold standard, defined as the result of expert manual lesion segmentations, is used instead. The difficulty to obtain reliable manual segmentations of lesions naturally limits the number of cases on which the method can be tested. The known inter- and intrarater variability in manual segmentation combined with the fact that radiologists may not agree in what to consider a lesion or not contributes further to the caution with which such human dependent validation should be considered. Instead of the result given by a unique expert, the combination of the segmentations realised by different operators may increase the trust with which to consider a gold standard.

The use of synthetic data can partially alleviate the problems related to man-

ually obtained gold standards. The Brainweb simulator (`http://www.http://brainweb.bic.mni.mcgill.ca/brainweb/`) provides on the same synthetic MR model the possibility to include MS lesions at mild, moderate and severe load. Within the simulator, variations in noise level and bias field effect can also be incorporated and therefore enables the validation of the algorithm under different conditions. Additionally, the availability of the simulator online enables for the comparison of the results across methods [100, 101]. Such a simulator has however some drawbacks: a single subject is represented thus bounding the statistical analysis, the images lack realism and the range of loads is quite limited. Alternatively, synthetic data can be created [103, 106, 156]. In those cases, masks of lesions are used to apply artificial damage on lesion-free scans [103]. Instead, additional lesion voxels can also be included in scans with existing lesions [156].

2.5.2 Measures of agreement

Numerous measures of agreement have been developed to assess the correspondences between an obtained segmentation and a reference delineation. Since each of them evaluates a specific characteristic of the segmentation quality, there is no one that stands out and gathers all the relevant information. This is the reason why it has been strongly suggested that such measurements should be combined and presented jointly when validating a new segmentation algorithm [157]. Generally those measures assess in different ways the relationships between the amount of true detection (true positives TP), erroneous detections (false positives FP) and missing voxels (false negatives FN). Due to the small amount of lesion voxels with respect to brain volumes some of the measures of agreement widely used in most generic segmentation validation framework are non optimal. Among the measures of agreement, Dice score coefficient (DSC=2TP/(2TP+FP+FN)), true positive rate or sensitivity (TPR=TP/(TP+FN)), false positive rate (FPR=FP/(TN+FP)), positive predictive value (PPV=TP/(TP+FP)), specificity (Spec=TN/(FP+TN)), and accuracy (Acc=(TP+TN)/(TP+TN+FP+FN)) are common measures of interest. Such assessments can be considered either per voxel or per lesion. Various other measures assessing for the difference in terms of shape based on the average surface may also be used. Section 5.2 will further detail and analyse these evaluation measures.

2.5.3 Surrogate measures of relevance

In complement to local measures of agreement, relevance of a lesion segmentation output can be assessed through the estimation of the correlation between obtained and reference total lesion loads. This can carry relevant information in terms of the bias of the evaluated algorithm with respect to the range of lesion load considered. Additionally, in order to assess the proficiency of an algorithm on a large dataset, correlations with clinical measures representative of the damage can be used as surrogate for relevance. For instance, Maillard et al. [158] validate their method on a large dataset by correlating the lesion load with the visual rating scores. A further way of validating a new method or at least gathering some insight on its performance is to try and compare it against existing previously validated methods. Comparisons on an in-house dataset may prove difficult if the methods tested are not publicly available and need to be reimplemented. Furthermore such comparisons may be subject to bias if the solutions target the specific cohort tested. To overcome these problems, in addition to simulator such as Brainweb described above, publicly available datasets with training and testing sets and gold standard manual segmentations, allow for a more direct comparison between methods. For instance, the data used for the MICCAI challenge 2008 on lesion segmentation [159] is nowadays widely used as part of the validation process.

Validation criteria are inherently dependent on the purpose of the lesion segmentation and methods are often developed to answer specific needs and problems. For instance, metrics related to the number of detected lesions will be more important in the case of MS for which the cardinality of lesions is assessed as treatment effect whereas it has been demonstrated not to be relevant in the case of age-related WMH [62]. Other criteria such as the bias of the method towards high or low lesion load or its clinical relevance are also characteristics that need to be assessed. Depending on the application, the sensitivity may be more important than the specificity. In other situations, for instance with elderly subjects that cannot lie still during scanning, robustness to noise and artefact may be specially important.

A complementary way to assess the validity of a segmentation technique is the evaluation of the results' reproducibility mentioned by García-Lorenzo [110]. Applying the tested methods to scans acquired in a very short time interval may provide a relevant measurement of potential bias. The lack of appropriate data make such a test

difficult in practice and performed in very few studies [49, 98, 103].

2.6 Generalisability

Across all the methods presented in this review, the question of generalisability is a major issue. In the case of methods using manual segmentation databases, this issue is raised as a strong drawback. For instance, in the work by Guizard et al. [112], a large part of the process success relies on the appropriate selection of subjects in the database that are similar enough to the tested data. One must however bear in mind that bias may appear in any design that attempts to tackle a specific problem. For instance, the tuning of extrinsic model parameters may be strongly biased by the set of examples chosen for comparison [89, 90, 120, 124]. In addition, initialisation frameworks, especially when relying on pathological subjects, introduce also some bias [98, 99, 121].

Therefore, testing performance in various conditions of scanners, protocols, resolution, preprocessing, modalities, range of disease severity are many parameters that require careful investigation. The question of the preprocessing is assessed by Steenwijk et al. [118] with respect to the intensity normalisation strategy, while many others have studied the influence of the choice of modalities on their results [81,91,93,97,112]. Although monospectral solutions are generally faster, do not risk coregistration errors and may be less demanding in term of MR protocol, multispectral information tend to yield better performance. Admiraal-Behloul et al. [49] warn however against a disproportionate amount of information that may potentially increase the results uncertainty. This effect is shown by Guizard et al. [112] for which the performance with T2 and FLAIR images is slightly higher than when including T1 modality to the feature space.

In many of the studies, the performance with respect to lesion load is evaluated, separating the subjects in groups of lesion loads (mild, moderate and severe). In general, the segmentation performance increases with the severity of WMH load. This is due both to the sensitivity of the evaluation measures to the load, with the noticeable case of the DSC, combined with the inherent difficulty of the detection at lower load that often correspond to cases in which the lesions are also much less prominent in term of signal. Yoo et al. [105] for instance found a positive correlation between lesion load and mean lesion intensity. In all methods, the fact that lesions represent a very small proportion of the total brain volume causes in fact problems in the balance of

the volumes when learning from databases since the amount of training samples from healthy tissues is much higher than the one from lesions. For instance, Deshpande et al. [144] addressed this issue with respect to the size of the dictionaries to be learnt for each of the classes. This is also the reason that leads Damangir et al. [140] to build a cascade of SVM in their segmentation process. With generative solutions modelling the lesions in a defined class, appearance variability and limited amount of samples have been warned as potential limitations to the robustness of the algorithm. Wango et al. [121] attempt to avoid the problem by limiting their space of interest for final lesion detection to the mask of WM and DGM. In contrast, methods that discard the lesions and other outliers based on trimming parameter may also lack in robustness for the estimation of the healthy parameters if the trimming is too high as experimented by García-Lorenzo [101]. In the work by Wang et al. [124], the trimming parameter, obtained via optimisation with respect to a database of manual segmentations was actually shown to be dependent on the lesion load and the performance differed accordingly. By contrast, in extreme cases of WMH lesion loads, such trimming may actually lead to an inverted segmentation between healthy and lesioned tissue.

A complementary subject of interest is the applicability of the method not only to different pulse sequences [86, 97, 100, 101, 123, 153] but also to different resolutions in case for instance of retrospective analysis. Schmidt et al. [90] performed such an analysis by artificially thickening the slice thickness of their images. Indeed, with thick slices, as may happen with FLAIR images, both partial volume and interpolation induced blurring are stronger. It has also been observed that 3D acquired FLAIR images were less affected by flow artefacts and tended to be more sensitive to lesions [160,161].

Chapter 3

Theoretical background

3.1 Gaussian mixture model and Expectation-Maximisation algorithm

In this chapter, the underlying background behind the Gaussian Mixture Model (GMM), the Expectation-Maximisation (EM) algorithm, initially introduced by Dempster et al. [147], and the possible variations that have been designed in the field of medical image segmentation will be described. The derivation of the EM algorithm follows the work by McLachlan [162]. Denoting $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$, the set of log-transformed normalised intensities indexed by n, with N the number of observations in the image, each voxel based $\mathbf{y}_n$ feature vector contains D intensity values, each of them associated to the channel d used for the segmentation. In Gaussian mixture models (GMM), the observed intensity is considered as originating in different mixing proportions from K Gaussian density distribution functions. Indexed by k, these Gaussian instances $\mathscr{G}(\mathbf{Y}|\theta_k)$, depend on the parameters $\theta_k = \{\boldsymbol{\mu}_k, \Lambda_k\}$, with $\boldsymbol{\mu}_k$ and Λ_k being respectively the mean and covariance matrix of the k^{th} component of the mixture, such that

$$\mathscr{G}(\mathbf{y}_n|\theta_k) = \frac{1}{(2\pi)^{D/2}|\Lambda_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}_n - \boldsymbol{\mu}_k)\Lambda_k^{-1}(\mathbf{y}_n - \boldsymbol{\mu}_k)^T\right).$$

Given the set of parameters $\boldsymbol{\Xi}_K$, the density for the mixed model in $\mathbf{y}_n$ is expressed as

$$f(\mathbf{y}_n|\boldsymbol{\Xi}_K) = \sum_{k=1}^{K} \omega_k \mathscr{G}(\mathbf{y}_n|\theta_k).$$

Under the constraint that the mixing coefficients ω_k of the mixture must be positive and sum to 1, and denoting $\boldsymbol{\omega}_K = \{\omega_1, \cdots, \omega_K\}$, the set of parameters $\boldsymbol{\Xi}_K = \{\Theta_K, \boldsymbol{\omega}\}$ where $\Theta_K = \{\theta_1, \cdots, \theta_K\}$ are optimised in order to best model the intensity distributions of the image and consequently provide an accurate labelling of the image.

The unknown labelling of the image, $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ is defined as hidden variables where $\mathbf{z}_n$, supporting label ℓ, is defined as the $\mathbf{e}_\ell$ vector of the canonical basis, *i.e.* the unity vector with component ℓ equal to 1 and all the others to 0. Introducing the complete data $\mathbf{X} = \{\mathbf{Y}, \mathbf{Z}\}$, the Expectation-Maximisation algorithm is used to optimise the data modelling and the classification at each voxel. The goal is then to maximise the marginal log-likelihood $\mathscr{L}(\mathbf{Y} : \boldsymbol{\Xi}_K) = \log(f(\mathbf{Y}|\boldsymbol{\Xi}_K))$. Since the distribution of the complete data is unknown, the Expectation-Maximisation enables a progressive increase of the marginal log-likelihood, by alternating between the estimation of the expectation of the complete data conditioned on the parameters (E-step), followed by its optimisation with respect to the parameters (M-step) using the decomposition

$$f(\mathbf{X} \mid \boldsymbol{\Xi}_K) = f(\mathbf{Y} \mid \boldsymbol{\Xi}_K) f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\Xi}_K).$$

At iteration t in the update of the parameters the marginal log-likelihood can be expressed as

$$\begin{aligned}
\log(f(\mathbf{Y}|\boldsymbol{\Xi}_K)) &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log(f(\mathbf{Y}|\boldsymbol{\Xi}_K))\right] \\
&= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(\frac{f(\mathbf{X}|\boldsymbol{\Xi}_K)}{f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K)}\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log(f(\mathbf{X}|\boldsymbol{\Xi}_K))\right] - \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log(f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K))\right] \\
&= \mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right) - \mathscr{H}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right),
\end{aligned}$$

where t represents the iteration number, $\mathbb{E}$ the expectation and

$$\begin{aligned}
\mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right) &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log(f(\mathbf{X}|\boldsymbol{\Xi}_K))\right] \\
\mathscr{H}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right) &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log(f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K))\right].
\end{aligned}$$

The concave nature of the log function enables the use of the Jensen inequality so that

$$\mathbb{E}\left[\log\left(\mathbf{X}\right)\right] \leqslant \log\left(\mathbb{E}\left[\mathbf{X}\right]\right).$$

Focusing on $\mathscr{H}$, it can be observed that:

$$\begin{aligned}\mathscr{H}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right) - \mathscr{H}\left(\boldsymbol{\Xi}_K^{(t)}\middle|\boldsymbol{\Xi}_K^{(t)}\right) =& \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K\right)\right)\right] \\ &- \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{Z}\middle|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}\right)\right)\right] \\ =& \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(\frac{f\left(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K\right)}{f\left(\mathbf{Z}\middle|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}\right)}\right)\right].\end{aligned}$$

Then, applying the Jensen equality expressed earlier :

$$\begin{aligned}\mathscr{H}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right) - \mathscr{H}\left(\boldsymbol{\Xi}_K^{(t)}\middle|\boldsymbol{\Xi}_K^{(t)}\right) &\leqslant \log\left(\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\frac{f\left(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K\right)}{f\left(\mathbf{Z}\middle|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}\right)}\right]\right) \\ &= \log\int\frac{f\left(\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K\right)}{f\left(\mathbf{Z}\middle|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}\right)} f\left(\mathbf{Z}\middle|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}\right) d\mathbf{Z} \\ &= 0.\end{aligned}$$

Thus,

$$\forall\, \boldsymbol{\Xi}_K,\ \mathscr{H}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right) \leqslant \mathscr{H}\left(\boldsymbol{\Xi}_K^{(t)}\middle|\boldsymbol{\Xi}_K^{(t)}\right).$$

This observation simplifies the Expectation-Maximisation algorithm since an increase in the $\mathscr{Q}$ function increases the marginal log-likelihood. Thus, the iterative process consists in alternating between:

Expectation step, in which $\mathscr{Q}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right)$ is estimated and

Maximisation step, which updates $\boldsymbol{\Xi}_K^{(t+1)} = \text{argmax}_{\boldsymbol{\Xi}_K}\mathscr{Q}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right)$.

Considering that the marginal log-likelihood is upper-bounded, the process will eventually converge.

3.1.1 E-step

Using the chain rule, the log-likelihood of the complete data can be expressed as

$$\begin{aligned}\log\left(f\left(\mathbf{Y},\mathbf{Z}|\boldsymbol{\Xi}_K\right)\right) &= \log\left(f\left(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Xi}_K\right)\cdot f\left(\mathbf{Z}|\boldsymbol{\Xi}_K\right)\right)\\ &= \log\left(f\left(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Xi}_K\right)\right)+\log\left(f\left(\mathbf{Z}|\boldsymbol{\Xi}_K\right)\right).\end{aligned}$$

Assuming independence between the N observations, it can be written as

$$\begin{aligned}\log\left(f\left(\mathbf{Y},\mathbf{Z}|\boldsymbol{\Xi}_K\right)\right) &= \log\left(\prod_n f\left(\mathbf{y}_n|\mathbf{z}_n,\boldsymbol{\Xi}_K\right)\right)+\log\left(\prod_n f\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right)\\ &= \sum_{n=1}^{N}\log\left(f\left(\mathbf{y}_n|\mathbf{z}_n,\boldsymbol{\Xi}_K\right)\right)+\sum_{n=1}^{N}\log\left(f\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right).\end{aligned}$$

Introducing $\mathbf{u}\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)$ and $\mathbf{s}\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)$ two vectors whose k^{th} components are respectively

$$u_k\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)=\log\left(f\left(\mathbf{y}_n|\mathbf{z}_n=\mathbf{e}_k,\boldsymbol{\Xi}_K\right)\right)$$

and

$$s_k\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)=\log\left(f\left(\mathbf{z}_n=\mathbf{e}_k|\boldsymbol{\Xi}_K\right)\right)=\log\left(\omega_k\right),$$

the expressions are simplified into

$$\begin{aligned}\log\left(f\left(\mathbf{y}_n|\mathbf{z}_n,\boldsymbol{\Xi}_K\right)\right) &= \mathbf{z}_n^T\mathbf{u}\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)\\ \log\left(f\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right) &= \mathbf{z}_n^T\mathbf{s}\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right).\end{aligned}$$

Replacing these in $\mathscr{Q}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right)$ leads to the following:

$$\begin{aligned}\mathscr{Q}\left(\boldsymbol{\Xi}_K\middle|\boldsymbol{\Xi}_K^{(t)}\right) &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{X}|\boldsymbol{\Xi}_K\right)\right)\right]\\ &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{Y}|\mathbf{Z},\boldsymbol{\Xi}_K\right)\right)\right]+\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{Z}|\boldsymbol{\Xi}_K\right)\right)\right]\\ &= \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\sum_{n=1}^{N}\log\left(f\left(\mathbf{y}_n|\mathbf{z}_n,\boldsymbol{\Xi}_K\right)\right)\right]+\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\sum_{n=1}^{N}\log\left(f\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right)\right]\\ &= \sum_{n=1}^{N}\mathbb{E}_{\mathbf{z}_n|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{y}_n|\mathbf{z}_n,\boldsymbol{\Xi}_K\right)\right)\right]+\sum_{n=1}^{N}\mathbb{E}_{\mathbf{z}_n|\mathbf{Y},\boldsymbol{\Xi}_K^{(t)}}\left[\log\left(f\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right)\right]\end{aligned}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n|\mathbf{y}_n,\boldsymbol{\Xi}_K^{(t)}} \left[\mathbf{z}_n^T \mathbf{u}\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right)\right] + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n|\mathbf{y}_n,\boldsymbol{\Xi}_K^{(t)}} \left[\mathbf{z}_n^T \mathbf{s}\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right)\right]$$
$$= \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n|\mathbf{y}_n,\boldsymbol{\Xi}_K^{(t)}} \left[\mathbf{z}_n^T\right] \mathbf{u}\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right) + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n|\mathbf{y}_n,\boldsymbol{\Xi}_K^{(t)}} \left[\mathbf{z}_n^T\right] \mathbf{s}\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right).$$

The vector $\mathbf{p}_n^{(t+1)} = \mathbb{E}_{\mathbf{z}_n|\mathbf{y}_n,\boldsymbol{\Xi}_K^{(t)}} \left[\mathbf{z}_n^T\right]$ can be seen as the vector gathering the normalised responsibilities

$$p_{nk}^{(t+1)} = f\left(\mathbf{z}_n = \mathbf{e}_k \middle| \mathbf{y}_n, \boldsymbol{\Xi}_K^{(t)}\right).$$

Applying the Bayes' Rule then leads to

$$p_{nk}^{(t+1)} = \frac{f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_k, \boldsymbol{\Xi}_K^{(t)}\right) f\left(\mathbf{z}_n = \mathbf{e}_k \middle| \boldsymbol{\Xi}_K^{(t)}\right)}{\sum_{k'=1}^{K} f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_{k'}, \boldsymbol{\Xi}_K^{(t)}\right) f\left(\mathbf{z}_n = \mathbf{e}_{k'} \middle| \boldsymbol{\Xi}_K^{(t)}\right)}.$$

3.1.2 M-step

Making use of the previously introduced notations, the $\mathscr{Q}$ function is optimised with respect to the parameters by

$$\begin{aligned}
\mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right) &= \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} u_k\left(\mathbf{y}_n|\boldsymbol{\Xi}_K\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} s_k\left(\mathbf{z}_n|\boldsymbol{\Xi}_K\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(f\left(\mathbf{y}_n|\mathbf{z}_n = \mathbf{e}_k, \boldsymbol{\Xi}_K\right)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\omega_k\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\frac{1}{(2\pi)^{D/2}\left|\Lambda_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)\Lambda_k^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)^T\right)\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\omega_k\right).
\end{aligned}$$

To facilitate the optimisation, a variant of the EM, called Expectation Conditional Maximisation (ECM) is used here. It consists in the separate and successive optimisation of the parameters. Even though the $\mathscr{Q}$ function is not globally maximised, it increases at each step in any case. The convergence properties of the algorithm are therefore not affected. The independent updates of the Gaussian parameters are estimated as the value of the critical points with respect to the corresponding variables, since it can be shown that these critical points correspond to a maximum of the function. The update of $\boldsymbol{\mu}_k$ is for instance obtained as the solution of

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right)\Big|_{\boldsymbol{\mu}_k^{(t+1)}} = 0$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} p_{nk}^{(t+1)} \log \left(\frac{1}{(2\pi)^{D/2} \left|\Lambda_k^{(t)}\right|^{1/2}} \exp\left(-\frac{1}{2} \left(\mathbf{y}_n - \boldsymbol{\mu}_k\right) \Lambda_k^{-1(t)} \left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)^T \right) \right) \Bigg|_{\boldsymbol{\mu}_k^{(t+1)}} = 0$$

$$\sum_{n=1}^{N} p_{nk}^{(t+1)} \left(\mathbf{y}_n - \boldsymbol{\mu}_k\right) \Lambda_k^{-1(t)} \Bigg|_{\boldsymbol{\mu}_k^{(t+1)}} = 0.$$

The update for $\boldsymbol{\mu}_k$ is then

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} p_{nk}^{(t+1)} \mathbf{y}_n}{\sum_{n=1}^{N} p_{nk}^{(t+1)}}.$$

Using the update obtained for $\boldsymbol{\mu}_k$ in the expectation conditional maximisation framework, the update of the covariance matrix Λ_k is performed in a similar manner by solving

$$\frac{\partial}{\partial \Lambda_k} \mathscr{Q}\left(\cdots, \boldsymbol{\mu}_k^{(t+1)}, \cdots, \Lambda_k, \cdots \middle| \boldsymbol{\Xi}_K^{(t)} \right) \Big|_{\Lambda_k^{(t+1)}} = 0$$

$$\frac{\partial}{\partial \Lambda_k} \sum_{n=1}^{N} p_{nk}^{(t+1)} \log \left(\frac{1}{(2\pi)^{D/2} \left|\Lambda_k\right|^{1/2}} \exp\left(-\frac{1}{2} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right) \Lambda_k^{-1} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)^T \right) \right) \Bigg|_{\Lambda_k^{(t+1)}} = 0$$

$$\sum_{n=1}^{N} p_{nk}^{(t+1)} \left(\frac{\partial}{\partial \Lambda_k} \log\left(\left|\Lambda_k\right|^{-1/2} \right) + \frac{\partial}{\partial \Lambda_k} \left(-\frac{1}{2} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right) \Lambda_k^{-1} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)^T \right) \right) \Bigg|_{\Lambda_k^{(t+1)}} = 0$$

$$\sum_{n=1}^{N} p_{nk}^{(t+1)} \left(-\Lambda_k^{-1} + \Lambda_k^{-1} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right) \Lambda_k^{-1} \right) \Bigg|_{\Lambda_k^{(t+1)}} = 0.$$

Thus the update of Λ_k is

$$\Lambda_k^{(t+1)} = \frac{\sum_{n=1}^{N} p_{nk}^{(t+1)} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t+1)}\right)}{\sum_{n=1}^{N} p_{nk}^{(t+1)}}$$

In this classical EM framework, the mixing coefficients ω_k are updated using the Lagrange multipliers method to enforce the constraint $\sum_{k=1}^{K} \omega_k = 1$. The updated mixing coefficients must be chosen so that they satisfy

$$\frac{\partial}{\partial \omega_k} \mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)} \right) + \lambda \left(1 - \sum_{k=1}^{K} \omega_k \right) \Bigg|_{\omega_k^{(t+1)}} = 0$$

$$\sum_{n=1}^{N} p_{nk}^{(t+1)} \frac{\partial}{\partial \omega_k} \log(\omega_k) + \frac{\partial}{\partial \omega_k} \lambda \left(1 - \sum_{k=1}^{K} \omega_k \right) \Bigg|_{\omega_k^{(t+1)}} = 0$$

$$\sum_{n=1}^{N} p_{nk}^{(t+1)} \frac{1}{\omega_k} - \lambda \Bigg|_{\omega_k^{(t+1)}} = 0.$$

And thus, the update of the mixing coefficients is

$$\omega_k^{(t+1)} = \frac{\sum_{n=1}^{N} p_{nk}^{(t+1)}}{\sum_{n=1}^{N} \sum_{k=1}^{K} p_{nk}^{(t+1)}} = \frac{\sum_{n=1}^{N} p_{nk}^{(t+1)}}{N}.$$

3.2 Theme and variations over the EM algorithm in medical imaging

3.2.1 Bias Field correction

Due to variations occurring in the main magnetic field during the process of image acquisition, a smoothly varying intensity inhomogeneity (IIH) may be observed in the images [131]. This artefact, also called bias field (BF), can be modelled as a multiplicative linear combination of basis functions. At the spatial position v_n, the noisy signal $\exp(\mathbf{y}_n^c)$ is modified into

$$\exp(\mathbf{y}_n) = \exp(\mathrm{BF}(v_n)) \cdot \exp(\mathbf{y}_n^c),$$

where the bias field can be expressed as

$$\mathrm{BF}(v_n) = \sum_{m=1}^{M} \mathbf{c}_m \chi_m(v_n),$$

where $\mathbf{c}_m$ is the vector of d linear coefficients (one for each modality) of the basis function χ_m. The appropriate coefficients, needed to model the bias field, can be obtained progressively within the EM framework, as presented by Van Leemput et al. [163]. In this solution, the bias field coefficients $C = \{\mathbf{c}_1, \cdots, \mathbf{c}_M\}$ are considered as parameters to optimise and the complete set of parameters $\boldsymbol{\Xi}_K$ is then $\{\Theta_K, \boldsymbol{\omega}_K, \mathbf{C}\}$. In order to facilitate the optimisation of such parameters, the intensities of the image are log-transformed, as mentioned in Section 3.1, which renders the bias field additive instead of multiplicative. Despite the log-transformation, the observed noise is still assumed to

be Gaussian [164] and the corrected intensity vector at voxel n is expressed as

$$\begin{aligned}\mathbf{y}_n^{\mathrm{c}} &= \mathbf{y}_n - \mathrm{BF}(v_n) \\ &= \mathbf{y}_n - \sum_{m=1}^{M} \mathbf{c}_m \chi_m(v_n).\end{aligned}$$

Considering the bias field to be independent between acquisitions and the covariances diagonal, each coefficient c_{dm} can be optimised separately. The $\mathscr{Q}$ function to be maximised with respect to the bias field coefficients is then

$$\begin{aligned}\mathscr{Q}\left(\Xi_K \middle| \Xi_K^{(t)}\right) = &\sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\frac{1}{(2\pi)^{D/2}\left|\Lambda_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{y}_n^{\mathrm{c}} - \boldsymbol{\mu}_k\right)\Lambda^{-1}\left(\mathbf{y}_n^{\mathrm{c}} - \boldsymbol{\mu}_k\right)^T\right)\right) \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\omega_k\right).\end{aligned}$$

The first derivative along c_{md} in the expectation conditional maximisation (ECM) framework is

$$\frac{\partial}{\partial c_{md}} \mathscr{Q}\left(\Xi_K \middle| \Xi_K^{(t)}\right) = \sum_{n=1}^{N} \chi_m\left(v_n\right) \sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2}\left(y_{nd} - \mu_{kd} - \sum_{m'=1}^{M} c_{m'd}\chi_m\left(v_n\right)\right),$$

where σ_{kd}^2 is the value of the diagonal coefficient of the covariance matrix Λ_k for channel d. Nulling all the first partial derivative with respect to the bias field coefficients for channel d in $\mathbf{c}^{d(t+1)} = \left\{c_{1d}^{(t+1)}, \cdots, c_{Md}^{(t+1)}\right\}$ leads to the following vectorial expression:

$$\left[\begin{array}{c} \sum_{n=1}^{N} \chi_1\left(v_n\right) \sum_k \frac{p_{nk}^{(t+1)}}{\sigma_k^2}\left(y_{nd} - \mu_{kd}\right) \\ \vdots \\ \sum_{n=1}^{N} \chi_M\left(v_n\right) \sum_k \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2}\left(y_{nd} - \mu_{kd}\right) \end{array}\right] = \left[\begin{array}{c} \sum_{n=1}^{N} \chi_1\left(v_n\right) \sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2} \sum_{m'=1}^{M} c_{m'd}\chi_{m'}\left(v_n\right) \\ \vdots \\ \sum_{n=1}^{N} \chi_M\left(v_n\right) \sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2} \sum_{m'} c_{m'd}\chi_{m'}\left(v_n\right) \end{array}\right]\Bigg|_{\mathbf{c}^{d(t+1)}}.$$

Introducing τ_{nd} and $\bar{y}_{nd}$ such that

$$\begin{aligned}\tau_{nd} &= \sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2} \\ \bar{y}_{nd} &= \frac{\sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2}\mu_{kd}}{\sum_{k=1}^{K} \frac{p_{nk}^{(t+1)}}{\sigma_{kd}^2}},\end{aligned}$$

the previous equality constraint can be simplified in

$$\left[\begin{array}{c} \sum_{n=1}^{N} \chi_1(v_n)\,\tau_{nd}\,(y_{nd}-\bar{y}_{nd}) \\ \vdots \\ \sum_{n=1}^{N} \chi_M(v_n)\,\tau_{nd}\,(y_{nd}-\bar{y}_{nd}) \end{array}\right] = \left[\begin{array}{c} \sum_{n=1}^{N} \chi_1(v_n)\,\tau_{nd} \sum_{m=1}^{M} c_{m'd}\chi_{m'}(v_n) \\ \vdots \\ \sum_{n=1}^{N} \chi_M(v_n)\,\tau_{nd} \sum_{m'=1}^{M} c_{m'd}\chi_{m'}(v_n) \end{array}\right]\Bigg|_{\mathbf{c}^{d(t+1)}}.$$

Denoting the matrices $\boldsymbol{\chi}$, Υ_d and the vector $\mathbf{r}_d$ as

$$\boldsymbol{\chi} = \left[\begin{array}{ccc} \chi_1(v_1) & \cdots & \chi_1(v_N) \\ \vdots & \ddots & \vdots \\ \chi_M(v_1) & \cdots & \chi_M(v_N) \end{array}\right] \quad \Upsilon_d = diag\,(\tau_{nd}) \quad \mathbf{r}_d = \left[\begin{array}{c} y_{1d}-\bar{y}_{1d} \\ \vdots \\ y_{nd}-\bar{y}_{nd} \end{array}\right],$$

the update for $\mathbf{c}^{\mathbf{d(t+1)}}$ can thus be expressed as follows:

$$\boldsymbol{\chi}^T \Upsilon_d \mathbf{r}_d = \boldsymbol{\chi}^T \Upsilon_d \boldsymbol{\chi} \mathbf{c}^{d(t+1)}.$$

3.2.2 Spatial *a priori* knowledge

As underlined by Fraley et al. [165], the EM algorithm may suffer from a poor initialisation. One of the partial solutions proposed to tackle this problem has been to introduce *a priori* spatial anatomical information to the model through statistical atlases. The use of such atlases is quite common in the field of medical imaging and they represent at each location the *a priori* probability to belong to a certain tissue class [166]. Such information is usually obtained by the averaging of individual segmentations of a population. These priors can then be aligned to the image before the segmentation [163], or aligned during the EM optimisation [167]. When compared to the initial EM algorithm, the introduction of statistical atlases is equivalent to the transformation of the global *a priori* mixing weights ω_k into spatially varying mixing weights ω_{nk}. In this case, the mixing weights are not updated anymore and the normalised responsibilities become

$$p_{nk}^{(t+1)} = \frac{f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_k, \boldsymbol{\Xi}_K^{(t)}\right) \omega_{nk}}{\sum_{k'=1}^{K} f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_{k'}, \boldsymbol{\Xi}_K^{(t)}\right) \omega_{nk'}}.$$

However, as the populations used to build the atlases are not necessarily representative of the type of image being segmented, the validity of such *a priori* information has been challenged in the literature [168]. In fact, in the case of the commonly used ICBM atlas (`http://www.loni.usc.edu`), the scans were taken from healthy young volunteers and therefore may not be appropriate for an ageing population in which atrophy can be observed [49]. In order to account for this problem, atlases can be progressively adapted to a specific subject [168, 169]. This model assumes that the spatially varying mixing priors are derived from a Dirichlet distribution, noted $\mathscr{D}$, that is chosen as the natural conjugate prior to a labelling multinomial distribution. Using the˜diacritic mark to denote the adapted version of the spatially varying *a priori* mixing coefficients, the maximisation of the log-likelihood must take into account the prior distribution over the atlases. A new term $\log\left[f\left(\tilde{\mathbf{\Omega}}\right)\right]$ is added to the expression to maximise and the M-step is transformed in the optimisation of

$$\begin{aligned} f\left(\tilde{\mathbf{\Omega}}\right) &= \prod_{n=1}^{N} \mathscr{D}\left(\tilde{\omega}_n, \boldsymbol{\vartheta}_n\right) \\ &= \prod_{n=1}^{N} \frac{\prod_{k=1}^{K} \tilde{\omega}_{nk}^{(\vartheta_{nk}-1)}}{\mathscr{B}(\boldsymbol{\vartheta}_n)}, \end{aligned}$$

where $\mathscr{B}$ is a Beta function and $\boldsymbol{\vartheta}_n$ the vector of Dirichlet prior parameters for voxel n such that $\vartheta_{nj} = 1 + \varepsilon\omega_{nj}$, where ε is a positive parameter assessing the strength of the applied relaxation. The E-step is kept unchanged but the M-step consists now in the optimisation of

$$f\left(\mathbf{X}\middle|\mathbf{\Theta}_K, \tilde{\mathbf{\Omega}}, \mathbf{C}\right) \cdot f(\mathbf{\Theta}_K, \tilde{\mathbf{\Omega}}, \mathbf{C})$$

with respect to the parameters and the $\mathscr{Q}$ function becomes

$$\begin{aligned} \mathscr{Q}\left(\mathbf{\Xi}_K\middle|\mathbf{\Xi}_K^{(t)}\right) = &\sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\frac{1}{(2\pi)^{D/2}\left|\Lambda_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)\Lambda^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_k\right)^T\right)\right) \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \log\left(\tilde{\omega}_{nk}\right) + \log\left(\prod_{n=1}^{N} \frac{\prod_{k=1}^{K} \tilde{\omega}_{nk}^{(\vartheta_{nk}-1)}}{\mathscr{B}(\boldsymbol{\vartheta}_n)}\right). \end{aligned}$$

The Laplace multipliers method is used to solve the update of the parameters constraining the sum of the mixing coefficients to 1 at each voxel. At each voxel n, for each tissue

k, the maximisation process must then satisfy

$$\begin{aligned}
\frac{\partial}{\partial \tilde{\omega}_{nk}} \mathscr{Q}\left(\Xi_K \middle| \Xi_K^{(t)}\right) + \lambda \left(1 - \sum_{k=1}^{K} \tilde{\omega}_{nk}\right)\Bigg|_{\tilde{\omega}_{nk}^{(t+1)}} &= 0 \\
\frac{\partial}{\partial \tilde{\omega}_{nk}} \left(p_{nk}^{(t+1)} + \varepsilon \omega_{nk}\right) \log \tilde{\omega}_{nk} + \lambda \left(1 - \sum_{k=1}^{K} \tilde{\omega}_{nk}\right)\Bigg|_{\tilde{\omega}_{nk}^{(t+1)}} &= 0 \\
\frac{p_{nk}^{(t+1)} + \varepsilon \omega_{nk}}{\tilde{\omega}_{nk}^{(t+1)}} &= \lambda \\
\tilde{\omega}_{nk}^{(t+1)} &= \frac{p_{nk}^{(t+1)} + \varepsilon \omega_{nk}}{\lambda}.
\end{aligned}$$

Furthermore, the value chosen for λ is such that

$$\begin{aligned}
\lambda &= \sum_{k=1}^{K} \left(p_{nk} + \varepsilon \omega_{nk}\right) \\
&= 1 + \varepsilon
\end{aligned}$$

and consequently the final update is

$$\tilde{\omega}_{nk}^{(t+1)} = \frac{p_{nk}^{(t)} + \varepsilon \omega_{nk}}{\varepsilon + 1}.$$

However, as a consequence of the assumption of independence between observations, the responsibilities lack in smoothness. Therefore, the atlases obtained with such a direct update may not be as smooth as usual statistical atlases. To address this concern, a Gaussian kernel G_σ with standard deviation σ is applied by convolution to the responsibilities as a form of spatial regularization, similarly to previously described methods [168, 169]. Eventually, the update for the *a priori* mixing coefficients is

$$\tilde{\omega}_{nk}^{(t+1)} = (1 - \kappa)\omega_{nk} + \kappa(G_\sigma \star p_{nk}^{(t+1)}),$$

where $\kappa = \dfrac{1}{(\varepsilon + 1)}$ and $\star$ represents the convolution operator.

3.2.3 Spatial consistency through Markov Random Field

Again, as a result of the voxelwise independence assumption, labelling configurations may present a limited neighbourhood consistency. In order to promote such consis-

tency, the introduction of contextual constraints through Markov Random Fields (MRF) is a common solution. Using an energy-based approach, the added constrain consists in requiring a higher energy for unlikely neighbourhood configurations and make the labelling at one voxel dependent on the labeling of its neighbours. In the following, $\mathbf{N}_n$ denotes the 6-neighbourhood (east, west, north, south, top and bottom) of a given voxel n and $\mathbf{Z}_{\mathbf{N}_n} = \{\mathbf{z}_i\}_{i\in\mathbf{N}_n}$. In this approach, the density function of the labelling given the parameters is

$$f(\mathbf{Z}|\mathbf{\Omega}) \propto \prod_{n=1}^{N} \exp\left(-U_{\mathrm{MRF}}\left(\mathbf{z}_n|\mathbf{z}_{\mathbf{N}_n}\right)\right) \prod_{k=1}^{K} \omega_{nj}^{z_{nj}},$$
$$\text{where } U_{\mathrm{MRF}}\left(\mathbf{z}_n|\mathbf{z}_{\mathbf{N}_n}\right) = \mathbf{z}_n H \sum_{m\in\mathbf{N}_n} \mathbf{z}_m$$

and H is the matrix containing the energy value constraints. In order to account for the possible anisotropy of the data, the sum over the neighbourhood $\mathbf{N}_n$ can be weighted by the inverse of the distance between the central voxel n and its neighbour m. Given a matrix representation, different energy constraints can be applied to different tissue combinations. The main concern with this spatial consistency constrain lies in the intractability of the expectation step, that would require to take into account the whole space of possible configurations. To circumvent this computational difficulty, a well adopted approximation, known as the mean field approximation can be used. This mean field approximation consists in assuming that the dependence of the central value to all its surrounding can be approximated by its dependence to the mean of its nearest neighbours [170]. This approximation can be expressed as

$$\begin{aligned} f(\mathbf{z}_n|\mathbf{\Omega}) &\simeq \exp(-U_{\mathrm{MRF}}(\mathbf{z}_n|\mathbb{E}[\mathbf{Z}_{\mathbf{N}_n}])) \prod_{k=1}^{K} \omega_{nk}^{z_{nk}} \\ &= \exp(-U_{\mathrm{MRF}}(\mathbf{z}_n)|\mathbf{P}_{\mathbf{N}_n})) \prod_{k=1}^{K} \omega_{nk}^{z_{nk}} \\ f(\mathbf{z}_n = \mathbf{e}_k|\mathbf{\Omega}) &\simeq \frac{\exp(-U_{\mathrm{MRF}}(\mathbf{e}_k)|\mathbf{P}_{\mathbf{N}_n}))\omega_{nk}}{\sum_{k'=1}^{K} \exp\left(-U_{\mathrm{MRF}}\left(\mathbf{e}_{k'}|\mathbf{P}_{\mathbf{N}_n}\right)\right)\omega_{nk'}}, \end{aligned}$$

where $\mathbf{Z}_{\mathbf{N}_n} = \{\mathbf{z}_i\}_{i\in\mathbf{N}_n}$ and $\mathbf{P}_{\mathbf{N}_n} = \{\mathbf{p}_i\}_{i\in\mathbf{N}_n}$. Thus the normalised responsibilities are transformed into

$$p_{nk}^{(t+1)} = \frac{f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_k, \boldsymbol{\Xi}_K^{(t)}\right) \exp\left(-U_{\mathrm{MRF}}\left(\mathbf{e}_k \middle| \mathbf{P}_{\mathbf{N}_n}^{(t)}\right)\right) \omega_{nk}}{\sum_{k'=1}^{K} f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_{k'}, \boldsymbol{\Xi}_K^{(t)}\right) \exp\left(-U_{\mathrm{MRF}}\left(\mathbf{e}_{k'} | \mathbf{P}_{\mathbf{N}_n}^{(t)}\right)\right) \omega_{nk'}}.$$

3.2.4 Constraint over covariance matrix

Under the assumption that the noise model does not change across the different tissues, it can be enforced by constraining the covariance matrix of the Gaussian components. Assuming the initial intensities for a component are random variables such that

$$\exp(\mathbf{y}) = \mathrm{BF} \cdot (\exp(\boldsymbol{\mu}) + \mathfrak{e}),$$

where $\mathfrak{e}$ represents a Gaussian white noise, then the log-transformed intensities follow

$$\begin{aligned}\mathbf{y} &= \mathrm{BF} + \boldsymbol{\mu} + \log\left(1 + \frac{\mathfrak{e}}{\exp(\boldsymbol{\mu})}\right) \\ &\simeq \mathrm{BF} + \boldsymbol{\mu} + \frac{\mathfrak{e}}{\exp(\boldsymbol{\mu})}.\end{aligned}$$

Considering that $\mathfrak{e}$ follows the same Gaussian distribution in the whole image, for each tissue, then for all k, the prior distribution of Λ_k can be modelled as an Inverse-Wishart distribution $\mathscr{W}^{-1}$ of parameters $S_k \Psi S_k$ and N denoted $\mathscr{W}^{-1}_{S_k\Psi S_k, N}(\Lambda_k)$ where S_k is a diagonal scaling matrix such that $S_k(d,d) = \dfrac{1}{\exp(\mu_{kd})}$. Depending on the parameters Ψ and $\nu = N$, the prior distribution over Λ_k follows

$$\mathscr{W}^{-1}_{S_k\Psi S_k, N}(\Lambda_k) \propto \frac{\left|S_k \Psi S_k\right|^{\frac{N}{2}}}{\left|\Lambda_k\right|^{\frac{N+D+1}{2}}} \exp\left[-\frac{1}{2}\mathrm{Tr}\left(\Lambda_k^{-1} S_k \Psi S_k\right)\right],$$

where Tr refers to the trace of the matrix. The inclusion of such a constraint in the covariance matrix contributes to a change in the $\mathscr{Q}$ function [171] that becomes

$$\mathscr{Q}\left(\boldsymbol{\Xi}_K \middle| \boldsymbol{\Xi}_K^{(t)}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \mathbf{u}_k(\mathbf{y}_n | \boldsymbol{\Xi}_K) + \sum_{n=1}^{N}\sum_{k=1}^{K} p_{nk}^{(t+1)} \mathbf{s}_k(\mathbf{z}_n | \boldsymbol{\Xi}_K) + \log\left(\prod_{k=1}^{K} \mathscr{W}^{-1}_{S_k\Psi S_k, N}(\Lambda_k)\right).$$

When optimising first the constraint parameter Ψ, the differentiation leads to the following constraint:

$$\frac{\partial}{\partial \Psi} \sum_{k=1}^{K} \left[\log |\Psi|^{N/2} - \frac{1}{2} \mathrm{Tr}\left(S_k^{(t)} \Psi S_k^{(t)} \Lambda_k^{-1} \right) \right] \Bigg|_{\Psi^{(t)}} = 0$$

$$K N \Psi^{(t)-1} = \sum_{k=1}^{K} S_k^{(t)} \Lambda_k^{-1} S_k^{(t)}.$$

At this stage of the update, Λ_k can be approximated by

$$\Delta_k^{(t)} = \frac{\sum_{n=1}^{N} p_{nk}^{(t)} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t)} \right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t)} \right)}{\sum_{n=1}^{N} p_{nk}^{(t)}}$$

and the update of Ψ is thus

$$\Psi^{(t)-1} = \frac{\sum_{k=1}^{K} S_k^{(t)} \Delta_k^{(t)-1} S_k^{(t)}}{N \cdot K}.$$

The differentiation with respect to Λ_k for the update of the covariance matrix verifies

$$\frac{\partial}{\partial \Lambda_k} \sum_{n=1}^{N} \sum_{k=1}^{K} p_{nk}^{(t)} \mathbf{u}_k \left(\mathbf{y}_n | \Xi_K \right) = \frac{1}{2} \sum_{n=1}^{N} p_{nk}^{(t)} \left(\Lambda_k^{-1} \left(\mathbf{y}_n - \boldsymbol{\mu}_k \right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_k \right) \Lambda_k^{-1} - \Lambda_k^{-1} \right)$$

$$\frac{\partial}{\partial \Lambda_k} \log \left(\mathscr{W}_{S_k \Psi S_k, N}^{-1} \left(\Lambda_k \right) \right) = \frac{\partial}{\partial \Lambda_k} \log \left(|\Lambda_k|^{-(N+D+1)/2} \right) - \frac{1}{2} \mathrm{Tr}\left(S_k \Psi S_k \Lambda_k^{-1} \right)$$

$$= -\frac{N+D+1}{2} \Lambda_k^{-1} + \frac{1}{2} \Lambda_k^{-1} S_k \Psi S_k \Lambda_k^{-1}.$$

The updated covariance form is then derived to satisfy

$$\frac{1}{2} \left(\sum_{n=1}^{N} p_{nk}^{(t)} \left[\Lambda_k^{-1(t)} \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t)} \right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_k^{(t)} \right) \Lambda_k^{-1(t)} - \Lambda_k^{-1(t)} \right] \right)$$

$$- \frac{N+D+1}{2} \Lambda_k^{-1(t)} + \frac{1}{2} \Lambda_k^{-1(t)} S_k^{(t)} \Psi^{(t)} S_k^{(t)} \Lambda_k^{-1(t)} = 0$$

$$\Lambda_k^{-1(t)} \Delta_k^{(t)} \Lambda_k^{-1(t)} + \frac{\Lambda_k^{-1(t)} S_k^{(t)} \Psi^{(t)} S_k^{(t)} \Lambda_k^{-1(t)}}{\sum_{n=1}^{N} p_{nk}^{(t)}} = \Lambda_k^{-1(t)} + \frac{N+D+1}{\sum_{n=1}^{N} p_{nk}^{(t)}} \Lambda_k^{-1(t)}$$

$$\Lambda_k^{-1(t)} \sum_{n=1}^{N} p_{nk}^{(t)} \Delta_k^{(t)} + \Lambda_k^{-1(t)} S_k^{(t)} \Psi^{(t)} S_k^{(t)} = \sum_{n=1}^{N} p_{nk}^{(t)} + (N+D+1)$$

$$\Lambda_k^{(t+1)} = \frac{\Delta_k^{(t)} + \frac{S_k^{(t)} \Psi^{(t)} S_k^{(t)}}{\sum_{n=1}^{N} p_{nk}^{(t)}}}{1 + \frac{N+D+1}{\sum_{n=1}^{N} p_{nk}^{(t)}}}.$$

Since $\boldsymbol{\mu_k}$ is included is the definition of the Inverse Wishart distribution through the S_k matrix, the update for $\boldsymbol{\mu_k}$ should also be altered. For the univariate case for instance, the update of μ_k should then satisfy:

$$\sum_{n=1}^{N} p_{nk}^{(t)} \mu_k^{(t+1)} - \frac{\psi^{(t)}}{\Lambda_k{}^{(t)}} \exp(-2\mu_k^{(t+1)}) = \sum_{n=1}^{N} p_{nk}^{(t)} y_n - N$$

Assuming a slow change in the update of the parameters and considering the mode for Λ_k to be in $\frac{\exp(-2\mu_k)\Psi}{N+D+1}$, the variation due to the inclusion of the Inverse Wishart distribution is considered negligeable and the update of μ is therefore not altered.

3.3 Model selection and evolution

In the context of GMM, an important parameter is the number of Gaussian components used to model the data. In the case of the classical EM optimisation, this information has to be chosen beforehand but is crucial to avoid both under- and overfitting. Various strategies have been developed to select the appropriate number of components in such a model. Figure 3.1 displays a possible classification of such strategies. Firstly it separates stochastic and resampling strategies from deterministic options that are usually based on a penalisation function. The second level indicates if the model selection process is performed on a comparison basis after convergence of the algorithm (end process comparison) or if this is done within the algorithm itself in a evolutionary framework (Within evolution). Lastly, the third criterion, defines the type of evolution that is applied to the model in terms of number of Gaussian components. Four types can be defined:

Static: when the number of components is constant throughout the evolution. It is generally related to an end process comparison framework.

Increasing: when the number of components can only increase.

Decreasing: when the number of components can only decrease.

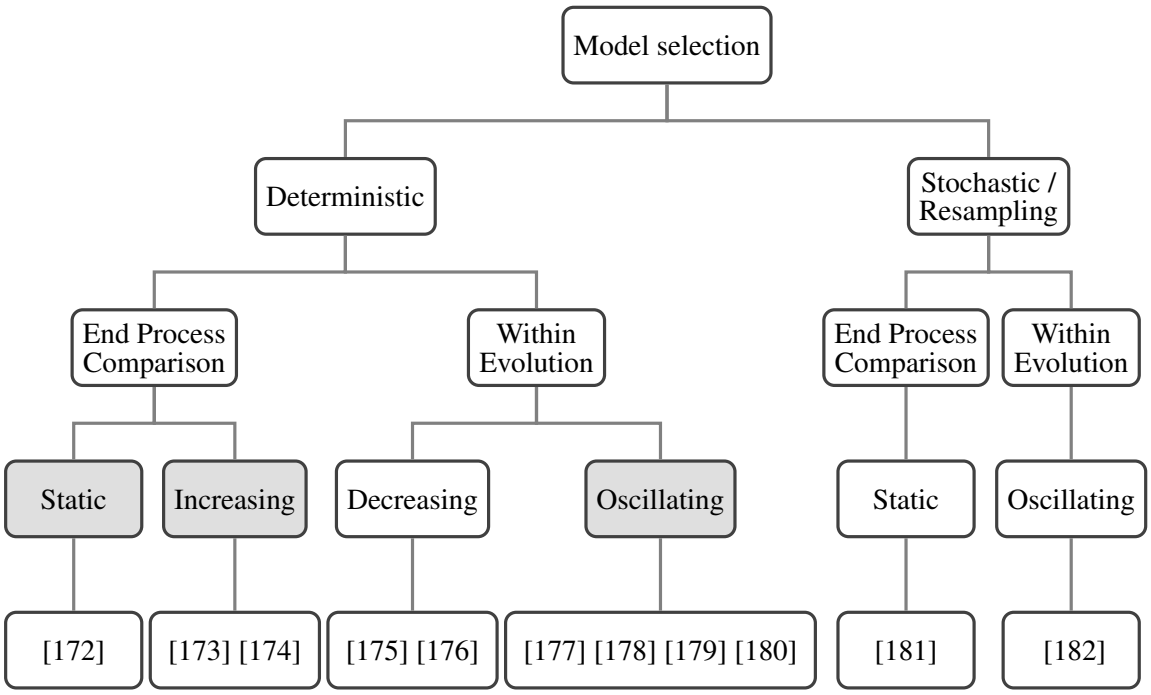


Figure 3.1: Classification of the different available strategies of model selection when considering GMM.

Oscillating: when the number of components can alternatively increase and decrease. Such algorithms generally do not require any more comparison with other models at convergence.

The shaded nodes in Figure 3.1 correspond to model selection processes for which split and merge strategies have been designed.

3.3.1 Examples of SM strategies for model selection

The split and merge (SM) strategies have not initially been introduced for the purpose of model selection but in order to deal with other drawbacks of the EM algorithm as documented by Fraley et al. [165]: the initialisation problem and the propensity of the EM to converge towards a local minimum [172]. Such problems arise from the non-optimal distribution of the Gaussian components over the data space. In the field of medical imaging, as developed in Section 3.2.2, when the Gaussian components can be matched with related anatomical tissues whose distribution is spatially well known, the initialisation problem can be partly solved through the introduction of anatomical

atlases. However, as soon as the model based on a unique Gaussian component per tissue is set aside, as with the solution of Ashburner et al. [167] where a fixed number of Gaussian components is assigned to each anatomical tissue, this issue may arise again. The split and merge strategies aim then to modify the distribution of the components by allowing Gaussian components for which the corresponding data observations are far from following a Gaussian distribution to be split in two and redundant Gaussian distribution pairs to be merged together. These modifications contribute to the redistribution of the components over the data space. In the static method using an end process comparison [172], at each iteration of the algorithm, a candidate triplet of Gaussian components is selected for a new model resulting in the split of the first component and the merge of the last two. The model is tested by running an EM until convergence on the proposed configuration. The log-likelihood of the model serves as objective function to decide whether to accept or not the tested model. The algorithms stops once all triplets configurations have been tested without meeting the acceptance criterion. This SM strategy designed for the reorganisation of the components over the data space can also be embedded in a model selection process. To that purpose, the algorithm is run over a range of number of components and the obtained models are tested against an objective function. The objective function is then the log-likelihood, penalised by a cost function that takes into account the complexity of the model. In the "increasing" framework using an SM strategy as presented by Blekas et al. [174], the split and the merge operations are decoupled: the split operation is always accepted and the subsequent tested merge operation is accepted only if the log-likelihood of the model increases. It evolves until reaching $K_{\max}$ Gaussian components in the model. To be used in a model selection perspective, this scheme requires a model comparison over a range of $K_{\max}$. Conversely, in the "oscillating" framework, the model selection process can be integrated within the framework, not requiring the comparison of a set of models at the end of the process. The model operations (split or merge) are decoupled, tested in an order defined in the method and checked for an increase in an objective function. A SM operation is generally tested after convergence of the EM [178, 180] but the SM operations can also be designed within the EM framework at each iteration [177]. The advantage of the split and merge strategies in the oscillating framework is that they allow for the simultaneous optimisation of the model parameters in a flexible manner

that may avoid local minima of the log-likelihood and the dynamic estimation of the appropriate number of model components.

3.3.2 Model evolution in SM strategies

For all the previously described SM strategies, three main problems related to the SM operations remain:

Initialisation problem How is a new model initialised after a split or a merge operation?

Precedence problem What is the precedence of the operations?

Acceptance criterion problem How to compare models and check for improvements?

Different solutions have been proposed in the literature to answer these three questions. The following subsections describe in more details the options adopted in the rest of this work.

3.3.2.1 Initialisation problem

In this subsection, a component with index k can be split into two Gaussian components k_1 and k_2. Respectively, for a merge operation, the components k_1 and k_2 can be merged into the component k. In order for a change to be tested, the new components of the model must be initialised. In order to preserve some initial information between the previous model and the newly tested configuration, an equality constraint over the first two probabilistic moments of the components is applied, as justified by Richardson et al. [182]. Such a constraint can be expressed as follows:

$$\begin{aligned}
\omega_k &= \omega_{k_1} + \omega_{k_2} \\
\omega_k \boldsymbol{\mu}_k &= \omega_{k_1} \boldsymbol{\mu}_{k_1} + \omega_{k_2} \boldsymbol{\mu}_{k_2} \\
\omega_k \left(\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T + \Lambda_k^T \Lambda_k \right) &= \omega_{k_1} \left(\boldsymbol{\mu}_{k_1}^T \boldsymbol{\mu}_{k_1} + \Lambda_{k_1} \Lambda_{k_1}^T \right) + \omega_{k_2} \left(\boldsymbol{\mu}_{k_2}^T \boldsymbol{\mu}_{k_2} + \Lambda_{k_2} \Lambda_{k_2}^T \right)
\end{aligned}$$

In the case of a merge operation, such a constraint can be solved straightforwardly, since the problem is well-posed. Following the solutions of Richardson and Zhang et al. [182, 183], the initial parameters for the component resulting of a merge operation

are

$$\begin{aligned}
\omega_k &= \omega_{k_1} + \omega_{k_2} \\
\boldsymbol{\mu}_k &= \frac{\omega_{k_1}\boldsymbol{\mu}_{k_1} + \omega_{k_2}\boldsymbol{\mu}_{k_2}}{\omega_{k_1} + \omega_{k_2}} \\
\Lambda_k &= \frac{\omega_{k_1}\left(\Lambda_{k_1} + \left(\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_k\right)^T\left(\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_k\right)\right) + \omega_{k_2}\left(\Lambda_{k_2} + \left(\boldsymbol{\mu}_{k_2} - \boldsymbol{\mu}_k\right)^T\left(\boldsymbol{\mu}_{k_2} - \boldsymbol{\mu}_k\right)\right)}{\omega_{k_1} + \omega_{k_2}}.
\end{aligned}$$

The two last equations can be justified with the work by Zhang et al. [183]:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n p\left(\mathbf{y}_n|\ell_n = k\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\frac{\omega_{k_1}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_1\right) + \mathbf{y}_n\frac{\omega_{k_2}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_2\right) \\
&= \frac{\omega_{k_1}\boldsymbol{\mu}_{k_1} + \omega_{k_2}\boldsymbol{\mu}_{k_2}}{\omega_k}.
\end{aligned}$$

In the same way, the value for the covariance matrix is obtained as

$$\begin{aligned}
\Lambda_k =& \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{y}_n p\left(\mathbf{y}_n|\ell_n = k\right) - \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k \\
=& \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{y}_n\frac{\omega_{k_1}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_1\right) + \mathbf{y}_n^T\mathbf{y}_n\frac{\omega_{k_2}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_2\right) - \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k \\
=& \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{y}_n\frac{\omega_{k_1}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_1\right) - \frac{\omega_{k_1}}{\omega_k}\boldsymbol{\mu}_{k_1}^T\boldsymbol{\mu}_{k_1} + \frac{\omega_{k_1}}{\omega_k}\boldsymbol{\mu}_{k_1}^T\boldsymbol{\mu}_{k_1} \\
&+ \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{y}_n\frac{\omega_{k_2}}{\omega_k}p\left(\mathbf{y}_n|\ell_n = k_2\right) - \frac{\omega_{k_2}}{\omega_k}\boldsymbol{\mu}_{k_2}^T\boldsymbol{\mu}_{k_2} + \frac{\omega_{k_2}}{\omega_k}\boldsymbol{\mu}_{k_2}^T\boldsymbol{\mu}_{k_2} \\
&- \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k \\
=& \frac{\omega_{k_1}\left(\Lambda_{k_1} + \boldsymbol{\mu}_{k_1}^T\boldsymbol{\mu}_{k_1}\right) + \omega_{k_2}\left(\Lambda_{k_2} + \boldsymbol{\mu}_{k_2}^T\boldsymbol{\mu}_{k_2}\right)}{\omega_k} - \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k.
\end{aligned}$$

As highlighted by Zhang et al. [183], these merging equations guarantee the new covariance matrix to be symmetric positive definite and allow the first and second moments not to be treated independently [172].

While the initialisation of the parameters when merging is well-posed, this is not the case for the splitting operation. Different methods have been proposed to solve this problem using for example the orthogonal decomposition of the covariance ma-

trix [183].

Symmetric positive definite, the covariance matrices can be decomposed as a product of orthogonal matrices under the following form:

$$\Lambda_k = Q_k^T Q_k = \sum_{d=1}^{D} \mathbf{q}_d^{(k)T} \mathbf{q}_d^{(k)}$$

$$\mathbf{q}_i^{(k)} \mathbf{q}_j^{(k)T} = \begin{cases} 0, & \text{if } i \neq j \\ \text{þ}_i^{(k)} & \text{otherwise} \end{cases},$$

where ðþ$^{(k)}$= $\left\{\text{þ}_1^{(k)}, \cdots, \text{þ}_D^{(k)}\right\}$ are the positive eigenvalues of the covariance matrix Λ_k ordered by decreasing value.

Relying on the work of Richardson [182] and Zhang et al. [183], the initialisation of the components results of a splitting operation is obtained as follows:

$$\begin{aligned}
\omega_{k_1} &= \omega_k \text{ð} & \omega_{k_2} &= \omega_k (1 - \text{ð}) \\
\boldsymbol{\mu}_{k_1} &= \boldsymbol{\mu}_k - \sqrt{\frac{\omega_{k_2}}{\omega_{k_1}}} \eta \mathbf{q}_1^{(k)} & \boldsymbol{\mu}_{k_2} &= \boldsymbol{\mu}_k + \sqrt{\frac{\omega_{k_1}}{\omega_{k_2}}} \eta \mathbf{q}_1^{(k)} \\
\Lambda_{k_1} &= \frac{\omega_{k_2}}{\omega_{k_1}} \Lambda_k + \frac{\omega_k}{\omega_{k_1}} \left(\xi - \xi \eta^2 - 1\right) \boldsymbol{q}_1^{(k)T} \boldsymbol{q}_1^{(k)} + \boldsymbol{q}_1^{(k)T} \mathbf{q}_1^{(k)} \\
\Lambda_{k_2} &= \frac{\omega_{k_1}}{\omega_{k_2}} \Lambda_k + \frac{\omega_k}{\omega_{k_2}} \left(\xi \eta^2 - \xi - \eta^2\right) \mathbf{q}_1^{(k)T} \mathbf{q}_1^{(k)} + \mathbf{q}_1^{(k)T} \mathbf{q}_1^{(k)}
\end{aligned} .$$

The free parameters ð,η, and ξ are all set to 0.5 similarly to what is chosen by Li et al. [178].

3.3.2.2 Operation precedence

In all the aforementioned SM strategies, a crucial aspect is the precedence of the operations. Intuitively it seems sensible to try to split a model component that does not fit correctly the corresponding observations whereas a merge operation may be more justified if it occurs between two distributions that are very similar. Again, various criteria have been described to sort the operations by likelihood of occurrence. Among them, the Kullback-Leibler divergence (KLD) or relative entropy, which measures how much information is lost when approximating a distribution P_1 by a distribution P_2, can not only be used for both the split and the merge operations but is also not biased by the relative weight of the different mixture components. Denoted $D_{KL}(P_1||P_2)$, the KLD is

generally defined as

$$D_{KL}\left(P_1||P_2\right) \equiv \sum_{\mathbf{y}} P_1\left(\mathbf{y}\right) \log\left(\frac{P_1\left(\mathbf{y}\right)}{P_2\left(\mathbf{y}\right)}\right).$$

The symmetric version of this expression, written as

$$\mathrm{KLD}\left(P_1||P_2\right) = D_{KL}\left(P_1||P_2\right) + D_{KL}\left(P_2||P_1\right),$$

will be used in the remaining of this work. In the splitting case, the KLD is defined between the model distribution $\mathscr{M}$ and the normalised frequency observations associated to the distribution. It measures how much information is lost when approximating the frequency distribution of the observations $p\left(\mathbf{Y},\theta_k\right)$ by the density function $\mathscr{M}$ with parameters θ_k, $\mathscr{M}\left(\mathbf{Y}|\theta_k\right)$. The normalised frequency distribution of the observations for component k, $p\left(\mathbf{Y}|\theta_k\right)$ is defined such that

$$p\left(\mathbf{y}|\theta_k\right) = \sum_{n=1}^{N} \frac{p_{nk}\delta\left(\mathbf{y}-\mathbf{y_n}\right)}{\sum_{n=1}^{N} p_{nk}},$$

where δ represents the Dirac distribution. The measure of KLD used for a split operation, $\mathrm{KLD_S}$ is defined as

$$\begin{aligned}\mathrm{KLD_S}(k|\Xi_K) &= D_{KL}\left(p\left(\mathbf{Y}|\theta_k\right)||\mathscr{M}\left(\mathbf{Y}|\theta_k\right)\right) + D_{KL}\left(\mathscr{M}\left(\mathbf{Y}|\theta_k\right)||p\left(\mathbf{Y}|\theta_k\right)\right) \\ &= \sum_{\mathbf{y}}\left[p\left(\mathbf{y}|\theta_k\right)\log\left(\frac{p\left(\mathbf{y}|\theta_k\right)}{\mathscr{M}\left(\mathbf{y}|\theta_k\right)}\right) + \mathscr{M}\left(\mathbf{y}|\theta_k\right)\log\left(\frac{\mathscr{M}\left(\mathbf{y}|\theta_k\right)}{p\left(\mathbf{y}|\theta_k\right)}\right)\right].\end{aligned}$$

For the merging operation the symmetric $\mathrm{KLD_M}$ measures how close the density distributions from two separated components are one from the other.

$$\begin{aligned}\mathrm{KLD_M}(k_1,k_2|\Xi_K) = &\sum_{n=1}^{N} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_1}\right)\log\left(\frac{\mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_1}\right)}{\mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_2}\right)}\right) \\ &+ \sum_{n=1}^{N} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_2}\right)\log\left(\frac{\mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_2}\right)}{\mathscr{M}\left(\mathbf{y}_n\middle|\theta_{k_1}\right)}\right).\end{aligned}$$

This criterion is commonly used to choose the order in which to test the different model changes.

3.3.2.3 Model acceptance criterion

The last problem to solve in an SM framework is the question of the acceptance criterion. Among the deterministic model selection frameworks making use of SM strategies, such a criterion is based on the maximisation of an objective function that combines a measure of the goodness of fit of the model to the observed data and a cost function penalising the complexity of the model. The penalty function is designed to avoid overfitting, while for the goodness of fit, the log-likelihood is the natural choice. Since the problem has been raised in the context of model selection by Akaike [184], many criteria have been developed to promote such a balance. Among those model selection criteria, the widely used Bayesian Inference Criterion (BIC) has been shown to perform well for GMM [185] and its derivation, following [186] and using the Bayesian-Laplace approximation is detailed hereafter. The BIC objective function is

$$f(K|\mathbf{Y}) \propto f(\mathbf{Y},K),$$

where, assuming that the priors on the parameters given the model are flat enough around the estimator of the set of parameters $\widehat{\boldsymbol{\Xi}_K}$ and the prior on the number of components is also flat, the likelihood can be expressed as follows:

$$\begin{aligned}
f(\mathbf{Y},K) &= f(K) f(\mathbf{Y}|K) \\
&= f(K) \int_{\boldsymbol{\Xi}_K} f(\boldsymbol{\Xi}_K|K)\, f(\mathbf{Y}|\boldsymbol{\Xi}_K)\, d\boldsymbol{\Xi}_K \\
&\approx f(K)\, f\left(\widehat{\boldsymbol{\Xi}_K}\right) \int_{\boldsymbol{\Xi}_K} f(\mathbf{Y}|\boldsymbol{\Xi}_K)\, d\boldsymbol{\Xi}_K \\
&= f(K)\, f\left(\widehat{\boldsymbol{\Xi}_K}\right) \cdot \int_{\boldsymbol{\Xi}_K} \exp\left[\mathscr{L}(\mathbf{Y}:\boldsymbol{\Xi}_K)\right] d\boldsymbol{\Xi}_K \\
&\approx f(K)\, f\left(\widehat{\boldsymbol{\Xi}_K}\right) \exp\left[\mathscr{L}\left(\mathbf{Y}:\widehat{\boldsymbol{\Xi}_K}\right)\right] \\
&\quad \cdot \int_{\boldsymbol{\Xi}_K} \exp\left[-\frac{1}{2}\left(\boldsymbol{\Xi}_K - \widehat{\boldsymbol{\Xi}_K}\right)^T \mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\boldsymbol{\Xi}_K}\right)\left(\boldsymbol{\Xi}_K - \widehat{\boldsymbol{\Xi}_K}\right)\right] d\boldsymbol{\Xi}_K \\
&= f(K)\, f\left(\widehat{\boldsymbol{\Xi}_K}\right) \exp\left[\log f\left(\mathbf{Y},\widehat{\boldsymbol{\Xi}_K}\right)\right] \\
&\quad \cdot \int_{\boldsymbol{\Xi}_K} \exp\left[-\frac{1}{2}\left(\boldsymbol{\Xi}_K - \widehat{\boldsymbol{\Xi}_K}\right)^T \mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\boldsymbol{\Xi}_K}\right)\left(\boldsymbol{\Xi}_K - \widehat{\boldsymbol{\Xi}_K}\right)\right] d\boldsymbol{\Xi}_K \\
&= f(K)\, f\left(\widehat{\boldsymbol{\Xi}_K}\right) f\left(\mathbf{Y}\middle|\widehat{\boldsymbol{\Xi}_K}\right) \frac{(2\pi)^{F(K)/2}}{\left|\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\boldsymbol{\Xi}_K}\right)\right|^{1/2}}.
\end{aligned}$$

$\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right) = \left[-\frac{\partial^2}{\partial\vartheta_m\partial\vartheta_l}\mathscr{L}\left(\mathbf{Y}:\mathbf{\Xi}_K\right)\right]\Big|_{\mathbf{\Xi}_K=\widehat{\mathbf{\Xi}_K}}$ is the Fisher information matrix of size $F(K)\times F(K)$ where $F(K)$ is the number of free parameters ϑ to determine and

$$\mathscr{L}\left(\mathbf{Y}:\mathbf{\Xi}_K\right) = \sum_{n=1}^{N}\log f\left(\mathbf{y}_n|\mathbf{\Xi}_K\right)$$
$$\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right) = \sum_{n=1}^{N}\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right).$$

Applying the law of large numbers, $\frac{1}{N}\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)$ can be approximated as $\mathbb{E}_{f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)}\left[\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right)\right]$. This approximation is further used to derive $\left|\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)\right|$ with $I_{F(K)}$ the identity matrix of dimension $F(K)$, number of free parameters in the mixture:

$$\begin{aligned}\left|\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)\right| &\approx \left|N\cdot\mathbb{E}_{f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)}\left[\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right)\right]\right| \\ &= \left|N\cdot I_{F(K)}\mathbb{E}_{f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)}\left[\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right)\right]\right| \\ &= N^{F(K)}\cdot\left|\mathbb{E}_{f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)}\left[\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right)\right]\right|.\end{aligned}$$

Using the above approximations and applying the log to the likelihood results in

$$\begin{aligned}\log f\left(\mathbf{Y},K\right) &\propto \log\left[f\left(K\right)f\left(\widehat{\mathbf{\Xi}_K}\right)f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)\frac{(2\pi)^{F(K)/2}}{\left|\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)\right|^{1/2}}\right] \\ &= \log f\left(K\right)+\log f\left(\widehat{\mathbf{\Xi}_K}\right)+\mathscr{L}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right) \\ &\quad+\frac{F(K)}{2}\log 2\pi-\frac{1}{2}\log\left|\mathbf{I}_{\mathscr{L}}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)\right| \\ &\approx \log f\left(K\right)+\log f\left(\widehat{\mathbf{\Xi}_K}\right)+\mathscr{L}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right) \\ &\quad+\frac{F(K)}{2}\log 2\pi-\frac{F(K)}{2}\log N-\frac{1}{2}\log\left|\mathbb{E}_{f\left(\mathbf{Y}\middle|\widehat{\mathbf{\Xi}_K}\right)}\left[\mathbf{I}_{\mathscr{L}}\left(\mathbf{y}_n:\widehat{\mathbf{\Xi}_K}\right)\right]\right|.\end{aligned}$$

When $N\to\infty$ only the terms function of N remain and the final expression is obtained:

$$\log f\left(\mathbf{Y},K\right)\approx\mathscr{L}\left(\mathbf{Y}:\widehat{\mathbf{\Xi}_K}\right)-\mathscr{P}\left(K\right)=\mathrm{BIC}(K).$$
$$\text{with }\mathscr{P}\left(K\right)=\frac{F(K)}{2}\log N.$$

In the case of a GMM using K Gaussian components in D modalities, the number of free parameters in the expression of the log-likelihood is

$$\begin{aligned} F(K) &= \underbrace{D \cdot K}_{\text{means}} + \underbrace{K \cdot \frac{D \cdot (D+1)}{2}}_{\text{covariance}} + \underbrace{K-1}_{\text{mixing coeffs.}} \\ &= K \cdot \frac{D^2+3D+2}{2} - 1 \end{aligned}$$

The value of $K-1$ for the mixing coefficient takes into account the constraint they sum to 1.

Decimation of the number of independent elements: Assuming independence between the observations enables the direct calculation of the complete log-likelihood, but this strong assumption should be considered carefully. Applying the BIC in such conditions would indeed be completely meaningless for medical images in which the number of voxels is extremely large (10^6). As proposed by Worsley et al. [187, 188] and applied by Groves et al. [189], a correction can be applied to this assumption by using a decimating factor that accounts for the proportion of observations that are actually independent. This decimating factor has been shown to be approximated by

$$\nu = \left(\frac{4\log 2}{\pi}\right)^{3/2} \prod_{r\in\{x,y,z\}} \frac{1}{\text{FWMH}_r}$$
$$\text{where } \text{FWMH}_r^2 = -\frac{2\log 2}{\text{corr}_r},$$

with corr_r the correlation between adjacent voxels in the r direction. Considering multispectral vectorial images, the correlation for the multispectral dataset is estimated as the mean of the correlations calculated for the D channels [190] using directly the log-transformed normalised data corrected for bias field. The corrected BIC can thus be expressed as:

$$\text{BIC}(K) = \nu\mathscr{L}\left(\mathbf{Y} : \widehat{\boldsymbol{\Xi}_K}\right) - \frac{F(K)}{2}\log(\nu N).$$

Chapter 4

Model selection with split and merge strategies

In this chapter, the tools described in previous chapters are assembled in order to build a hierarchical adaptive split and merge GMM model selection framework. When building the presented schemes, the main motivation was to use the final model to distinguish damaged from healthy white matter as separated Gaussian components. A naive model and its limitations is first presented before detailing BaMoS (Bayesian Model Selection) with its needed refinements and implementation adjustments.

4.1 Naive model: BiASM

Naively, one may design an algorithm that encompasses these SM concepts and the anatomical prior information so as to define directly the number of Gaussian components necessary to model each tissue class appropriately. In this framework named BiASM, standing for **Bi**layered **A**natomically-constrained **S**plit and **M**erge expectation maximisation algorithm [191], J tissue classes, indexed by j, are modelled as a mixture of K_j Gaussian components, leading to a bilayered graphical model displayed in Figure 4.1 .

Within a brain segmentation application, the tissue classes are defined as the grey matter (GM), the white matter (WM), the cerebrospinal fluid (CSF) and the non-brain (NB) tissues (bone, fat, skin, muscle and others which remain after skull-stripping). The model is anatomically constrained using statistical atlases and an MRF constraint. Even though BiASM is able to separate different tissue types from lesions, it suffers from various limitations that are described hereafter:

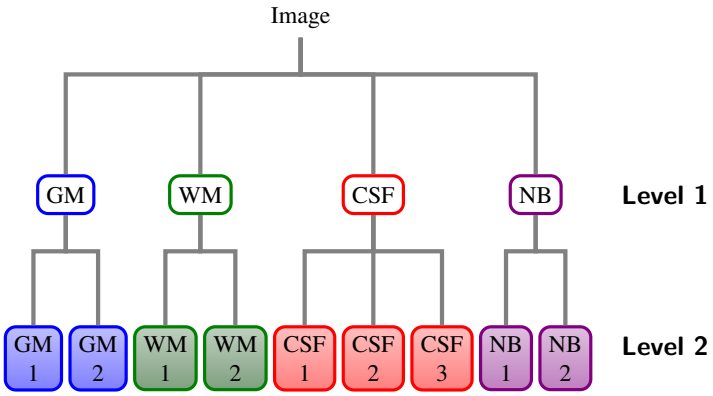


Figure 4.1: Example of the hierarchical description of the naive bilayered model.

Automatic selection Due to the variability of lesion location within the brain and the smoothness of the statistical atlases, the Gaussian components that could be interpreted as lesion-related are not solely associated to white matter. Without knowledge about what to consider as normal healthy tissue, an automatic selection of the lesion-related elements is therefore problematic.

Transfer problem The ability of the framework to add new Gaussian components increases the flexibility to the intensity modelling. Thus, the priors become more important in constraining the segmentation as a new Gaussian distribution can be created to accommodate a region with high *a priori* probability. In BiASM, the GM/CSF partial volume is regularly modelled as pure GM, leading to problematic results.

Bias Field Not accounting for the presence of atypical intensities may contribute to a poor modelling of the bias field effect.

4.1.1 Introducing the outlier modelling

In order to deal with the limitations of the naive model, the modelling of outliers can be further investigated. As evoked in Section 2.3.2, such methods are popular when applied to images with WMH. Indeed, the EM algorithm is known to have a breakdown point of 0 [148], meaning that the presence of only one unexpected observation (outlier) may affect considerably the final result. The breakdown occurs when a degenerate component with singular covariance matrix is assigned to the outlier element, the likelihood being then unbounded. To ensure an appropriate convergence of the EM in the presence of outliers, without directly modelling the outlier distribution, a solution

is to limit the impact of the outliers by rejecting them from the parameters optimisation using the Trimmed Log-Likelihood [97, 101, 124, 192]. These only consider the fraction of elements best in line with the current estimator for the parameters update. Another solution is to weigh the contribution of each sample to the updated parameters by introducing a measure of outlierness either for each component of the GMM [79] or globally [148]. This allows to downweight the influence of the outlier observations. As the distributions of these unexpected observations are not known and can be sparse, a popular way has been to model their presence by the addition of a uniform distribution ($\mathscr{U}$) [193] over the range of possible intensities:

$$f(\mathbf{y}_n|\boldsymbol{\Xi}) = \sum_{k=1}^{K} \omega_k \mathscr{G}(\mathbf{y}_n|\theta_k) + \omega_{K+1}\mathscr{U}$$

where ω_{K+1} is the weight attributed to the uniform distribution so that $\sum_{k=1}^{K+1} \omega_k = 1$. As stated by Schroeter et al. [148], the use of the uninformative uniform distribution helps tackling the problem of the number of GMM components. Combining the idea of a rejection class with a SM strategy allows for modelling the different types of outliers and separating hyperintense lesions in the white matter from hypointense iron deposition that are linked in many neurodegenerative disorders [194].

4.1.1.1 Mahalanobis distance

As mentioned in Section 1.3.3, the choice of a threshold measuring the outlierness of an observation can be obtained through the use of the Mahalanobis distance. It is possible to draw a link between this measure and the probability for such observations to occur. When handling normally distributed multidimensional data of parameters $\boldsymbol{\mu}, \Lambda$, the Mahalanobis distance d_{Mahal} is often used to measure how far an observation vector $\mathbf{y}$ is from the corresponding mean of the distribution and defined as

$$\begin{aligned} d_{\text{Mahal}}(\mathbf{y}, \boldsymbol{\mu}, \Lambda) &= \| \mathbf{y} - \boldsymbol{\mu} \|_{\Lambda} \\ &= \sqrt{(\mathbf{y} - \boldsymbol{\mu}) \Lambda^{-1} (\mathbf{y} - \boldsymbol{\mu})^T}. \end{aligned}$$

The value of 3 for the Mahalanobis distance has been widely used as threshold to consider observations as outliers [79]. Denoting τ_{Mahal} the Mahalanobis distance threshold value above which the observations $\mathbf{y}$ should be treated as outliers, according to [195]

the probability for such an observation can be derived as

$$\begin{aligned} P(\| \mathbf{y}-\boldsymbol{\mu} \|_{\Lambda} \geq \tau_{\text{Mahal}}) &= 1 - P(\| \mathbf{y}-\boldsymbol{\mu} \|_{\Lambda} \leq \tau_{\text{Mahal}}) \\ &= 1 - \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \int_{\mathbf{y}\in R} \exp\left(-\frac{1}{2} d_{\text{Mahal}}(\mathbf{y},\boldsymbol{\mu})^2\right) d\mathbf{y}, \end{aligned}$$

with $R = \mathbf{y} | d_{\text{Mahal}}(\mathbf{y},\boldsymbol{\mu}) \leq \tau_{\text{Mahal}}$. By the change of variable $\boldsymbol{\rho} = \Lambda^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$ and using the fact that the Jacobian of the transformation is $|\Lambda|^{1/2}$, the previous equation can be transformed into

$$\begin{aligned} P(\| \mathbf{y}-\boldsymbol{\mu} \|_{\Lambda} \geq \tau_{\text{Mahal}}) &= 1 - \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \int_{\|\boldsymbol{\rho}\| \leq \tau_{\text{Mahal}}} \exp\left(-\frac{1}{2} \| \boldsymbol{\rho} \|^2\right) |\Lambda|^{1/2} d\boldsymbol{\rho} \\ &= 1 - \frac{1}{(2\pi)^{D/2}} \int_{\|\boldsymbol{\rho}\| \leq \tau_{\text{Mahal}}} \exp\left(-\frac{1}{2} \| \boldsymbol{\rho} \|^2\right) d\boldsymbol{\rho}. \end{aligned}$$

Thanks to the rotational invariance of the integrand, denoting $r = \| \boldsymbol{\rho} \|$ with $S^{D-1}(r)$ the sphere of radius r in dimension D-1, dr and dA respectively the length and the area parts of the elementary volume $d\boldsymbol{\rho}$, and noting that $A_{D-1}(r) = r^{D-1} A_{D-1}(1)$ with $A_{D-1}(1) = \frac{2\pi^{D/2}}{\Gamma(D/2)}$the transformation into spherical coordinates leads to

$$\begin{aligned} P(\| \mathbf{y}-\boldsymbol{\mu} \|_{\Lambda} \geq \tau_{\text{Mahal}}) &= 1 - \frac{1}{(2\pi)^{D/2}} \int_0^{\tau_{\text{Mahal}}} \int_{S^{D-1}(r)} \exp\left(-r^2/2\right) dA dr \\ &= 1 - \frac{1}{(2\pi)^{D/2}} \int_0^{\tau_{\text{Mahal}}} \exp\left(-r^2/2\right) \int_{S^{D-1}(r)} dA dr \\ &= 1 - \frac{1}{(2\pi)^{D/2}} \int_0^{\tau_{\text{Mahal}}} \exp\left(-r^2/2\right) A_{D-1}(r) dr \\ &= 1 - \frac{A_{D-1}(1)}{(2\pi)^{D/2}} \int_0^{\tau_{\text{Mahal}}} \exp\left(-r^2/2\right) r^{D-1} dr \\ &= 1 - \frac{A_{D-1}(1) 2^{D/2-1}}{(2\pi)^{D/2}} \int_0^{\tau^2_{\text{Mahal}}/2} e^{-t} t^{\frac{D}{2}-1} dt \\ &= 1 - \frac{1}{\Gamma(D/2)} \int_0^{\tau^2_{\text{Mahal}}/2} e^{-t} t^{\frac{D}{2}-1} dt \end{aligned}$$

by the change of variable $t = r^2/2$. Eventually the previous equation can be simplified into

$$P(\| \mathbf{y}-\boldsymbol{\mu} \|_{\Lambda} \geq \tau_{\text{Mahal}}) = 1 - \frac{\gamma(D/2, \tau^2_{\text{Mahal}}/2)}{\Gamma(D/2)},$$

Number of modalities	Mahalanobis distance value			
	2	2.5	3	3.5
1	0.0455	0.0124	0.0027	0.000465
2	0.1353	0.0439	0.0111	0.0022
3	0.2615	0.1001	0.02993	0.0066

Table 4.1: Probabilistic correspondence of Mahalanobis distance for different number of modalities

where the lower incomplete Gamma function is defined as $\gamma(\frac{D}{2}, \tau_{\text{Mahal}}) = \int_0^{\tau_{\text{Mahal}}} t^{\frac{D}{2}-1} e^{-t} dt$ and the Gamma function computed in $\frac{D}{2}$ is $\Gamma(D/2) = \int_0^{\infty} t^{\frac{D}{2}-1} e^{-t} dt$. It corresponds to the value of the cumulative distribution of the χ^2 law with D degrees of freedom taken in τ^2_{Mahal}. Table 4.1 presents the corresponding probabilistic values for different choices of Mahalanobis distance with varied number of modalities.

4.1.2 SM strategy with uniform distributions

In models that consider outliers under a uniform distribution as mentioned in Section 4.1.1, the application of a SM strategy requires the definition of a split of the uniform distribution. Many solutions such as transforming a uniform distribution into a Gaussian or splitting it into two Gaussian distributions can be proposed. As uniform distributions provide a trace probability to all observations, thus acting as a regulariser, they should be preserved in the model. The split of a uniform distribution is therefore defined as the transformation of a uniform distribution into one Gaussian and a remaining uniform distribution.

The next problem at hand is to define the parameters to initialise the newly formed Gaussian distribution. Using the mean of the observations corresponding to the uniform distributions is not truly meaningful since the presence of outliers at both extremes of the data spectrum would have a cancelling effect. As the choice of the initial covariance matrix for such a component is also problematic, a simple K-means algorithm can be used to separate the cluster corresponding to the uniform distribution into two subclasses. The solution of the K-means then provides an initial estimation of the mean and the covariance of the Gaussian component.

The problem is thus simplified into finding meaningful initial guesses for the means of the two clusters to build from the K-means algorithm. The two classes of the K-means are initialised using the mean and the mode of the distribution under con-

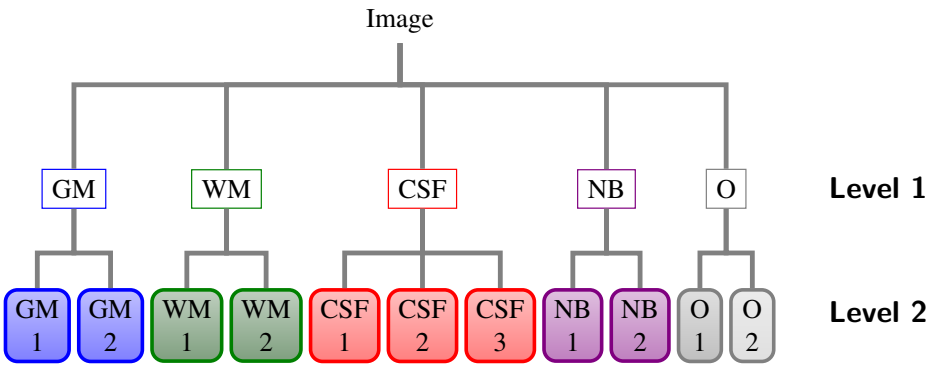


Figure 4.2: Example of the hierarchical description of BiASM modified for outlier detection in two distinctive layers. For the outlier mixture the node with a lighter shade on Level 2 correspond to the remaining uniform distribution.

sideration, where the mode is obtained through kernel density estimation. The initial parameters for the new Gaussian component are the mean and the covariance of the class with the lowest variance out of the two clusters obtained after running the K-means with the defined initial values. This educated guess of the initial estimation of the parameters of the Gaussian component resulting from the split of a uniform distribution is crucial to the evolution of the model and could be further improved and redefined in the future. An extra empirically-defined constraint prevents merges between uniform and Gaussian distributions. An example model is displayed in Figure 4.2, where the lighter shade in the outlier mixture correspond to the remaining uniform component.

4.1.3 Need for spatial separation for outliers

While from a purely theoretical point of view BiASM provides an interesting model of anatomy, several problems are present because of the intensity similarities between flow artefacts mostly present in the CSF and white matter lesions. Thus, in order to spatially separate outliers originating from different tissue classes, a three-level Gaussian Mixture model named BaMoS (Bayesian Model Selection) is now explored.

4.2 BaMoS

4.2.1 Class hierarchy

Considering an inlier/outlier differentiation on top of the anatomical tissue classification produces the following three-layer architecture [196]:

Level 1 At the first level (Level 1), indexed by l, the model is robustly divided into two density functions I and O, that correspond respectively to the inlier part (I), modelling the healthy tissues, and to the outlier part (O), related to the unexpected observations, such that

$$f(\mathbf{y}_n|\boldsymbol{\Xi}_\mathbf{K}) = b_I I(\mathbf{y}_n|\boldsymbol{\Xi}_\mathbf{K}) + b_O O(\mathbf{y}_n|\boldsymbol{\Xi}_\mathbf{K}),$$

with $b_I + b_O = 1$ and $b_l \geq 0$, and introducing $\mathbf{b}$ the vector formed with these parameters. Note that $O(\mathbf{y}_n|\boldsymbol{\Xi}_\mathbf{K})$ is a full mixture model, contrary to previously proposed models which assumed a uniform distribution for O [193].

Level 2 The second level (Level 2), indexed by j characterizes the anatomical tissue classes (*i.e.* if an inlier or outlier voxel belongs to WM, GM, CSF or other non-brain (NB) tissues). The number of anatomical classes J_l is considered the same for both the inlier and outlier classes since the model is built under an assumption of symmetry, simplifying $J_I = J_O = J$. The distribution is thus:

$$f(\mathbf{y}_n|\boldsymbol{\Xi}_\mathbf{K}) = \sum_{l\in I,O} b_l \sum_{j=1}^{J} a_{l_j} \Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right),$$

where b_l, a_{l_j} and $\Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right)$ are respectively the mixing weight of l, the class weight for l_j and the likelihood of the data at voxel n for the tissue class l_j. At this point, $\mathbf{a}$ denotes the vector of mixing weights a_{l_j} satisfying $\sum_{j=1}^{J} a_{l_j} = 1$ and $a_{l_j} \geq 0, \forall l \in \{I,O\}$.

Level 3 The third level (Level 3), indexed by k, characterizes the multiple intensity clusters of each inlier or outlier tissue class and models the acquisition noise in the observations from the expected biological mean signal. Each anatomical class density distribution is modelled by a mixture of multiple components with distribution $\mathscr{M}$, that can be Gaussian ($\mathscr{G}$) and/or uniform ($\mathscr{U}$) such that

$$\begin{aligned}\Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right) &= \sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right) \\ &= \sum_{k=1}^{K_{l_j}} w_{l_{j_k}} \mathscr{G}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right) + w_{l_{j_{K_{l_j}+1}}} \mathscr{U}_{l_j}\end{aligned}$$

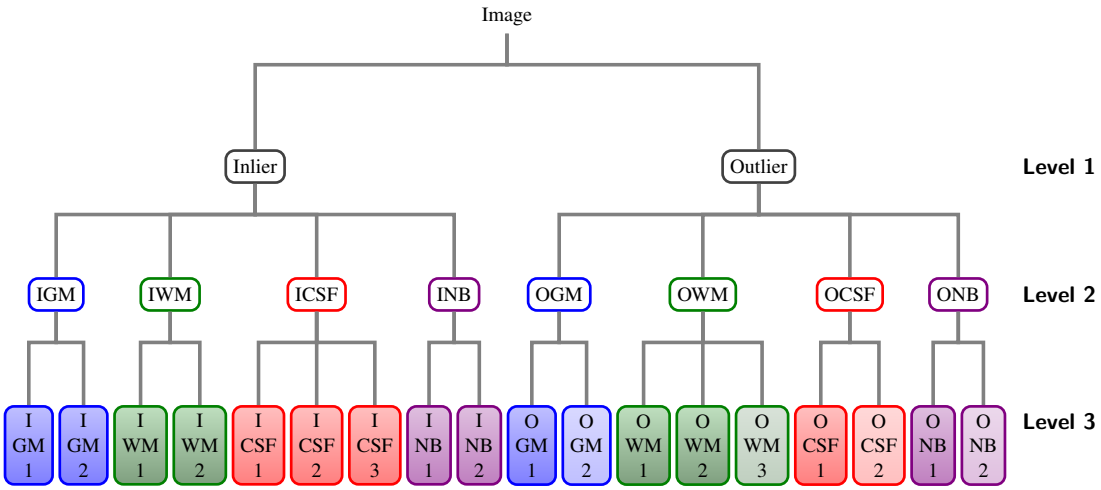


Figure 4.3: Example of the hierarchical description of BaMoS in three distinctive layers. Note that on the outlier part of the tree, the nodes with lighter shades on Level 3 correspond to remaining uniform distributions.

where K_{l_j} is the number of Gaussian components in class l_j, $w_{l_{j_k}}$ is the mixing proportion (≥ 0) of class l_{j_k} and $\theta_{l_{j_k}}$ are the corresponding Gaussian parameters. The uniform distribution in each class l_j is only parametrised by the mixing coefficient $w_{l\,j_{K_{l_j}+1}}$ with amplitude 1 over the range of intensities normalised between 0 and 1. The mixing coefficients for class l_j are gathered in the vector $\mathbf{w}_{l_j}$, with $\mathbf{W}$ being the set of all such vectors that satisfy $\sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} = 1$, $\forall l \in \{I,O\}$ and $\forall j \in \{1,\cdots,J\}$.

Adopting the notation $\omega_{l_{j_k}} = b_l a_{l_j} w_{l_{j_k}}$ with $\boldsymbol{\pi} = \{\mathbf{b},\mathbf{a},\mathbf{W}\}$ the set of *a priori* mixing weights at the different hierarchical levels and considering the observations as independent and identically distributed (*iid*), the multi-layered mixture model can finally be expressed as follows:

$$f(\mathbf{Y}|\boldsymbol{\Xi}_{\mathbf{K}}) = \prod_{i=1}^{N} \sum_{l\in I,O} \sum_{j=1}^{J} \left[\sum_{k=1}^{K_{l_j}+1} \omega_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n \middle| \theta_{l_{j_k}}\right) \right].$$

An example of a possible hierarchical model is displayed in Figure 4.3, where at Level 1 l takes the values I or O and j the values in {GM, WM, CSF, NB}.

In such a design, an atlas adaptation on both Level 1 and Level 2 is possible when adopting a symmetric modelling (symmetric in the J anatomical tissues modelled) providing at Level 1 new information on the possible location of the outliers. The problem mentioned in the case of BiASM is more easily solved due to the above differentiation

	BiASM	BaMoS
Number of hierarchical layers	2	3
Automatic selection of number of components	Yes	Yes
Atlas adaptation using Dirichlet priors	No	Yes
Automatic post-processing for pathology segmentation	No	Yes

Table 4.2: Characteristics comparison between BiASM and BaMoS

between inliers and outliers. As detailed later in Section 5.1 when applied to the lesion segmentation problem, this three layered model and especially the inlier/outlier separation enables an easier selection of the lesion-related elements based on comparison to the parameters used for the inlier-related GMM. Table 4.2 summarises the differences between the initial BiASM model and BaMoS.

4.2.2 EM extensions in BaMoS

4.2.2.1 Spatial knowledge and Dirichlet priors

The inclusion of statistical atlases to enforce some spatial knowledge is possible when the interpretation of the classes is *a priori* provided which is the case at Level 1 and Level 2. Thus, such anatomical atlases are introduced to inform on spatial location of the tissues at Level 2. At Level 1, the class mixing proportion is initially not spatially-variant. However, if a singular value is assigned to all voxels, such as in Section 3.2.2, then the mixing proportions can be modelled as a Dirichlet prior both at Level 1 and Level 2. Denoting $\mathbf{B}$ and $\mathbf{A}$ the atlases used at Level 1 and Level 2 respectively and $\tilde{\mathbf{\Omega}} = \left\{\tilde{\mathbf{B}}, \tilde{\mathbf{A}}, \mathbf{W}\right\}$, and noting the symmetry of the model between inliers and outliers, the prior distribution over the atlases becomes:

$$\begin{aligned} f\left(\tilde{\mathbf{\Omega}}\right) &= f\left(\tilde{\mathbf{B}}\right) f\left(\tilde{\mathbf{A}}\right) \\ &= \prod_{n=1}^{N} \mathscr{D}\left(\tilde{\mathbf{b}}_n, \boldsymbol{\beta}_n\right) \mathscr{D}\left(\tilde{\mathbf{a}}_n, \boldsymbol{\alpha}_n\right) \\ &= \prod_{n=1}^{N} \frac{\prod_{l \in I,O} \tilde{b}_{nl}^{(\beta_{nl}-1)}}{\mathscr{B}(\boldsymbol{\beta}_n)} \frac{\prod_{j=1}^{J} \tilde{a}_{nj}^{(\alpha_{nj}-1)}}{\mathscr{B}(\boldsymbol{\alpha}_n).} \end{aligned}$$

Adapting directly the solution derived in Section 3.2.2, the decoupled update of the atlases is solved as:

$$\tilde{b}_{nl}^{(t+1)} = \frac{\varepsilon_1 b_{nl} + p_{nl}^{(t+1)}}{\varepsilon_1 + 1} \qquad \tilde{a}_{nl}^{(t+1)} = \frac{\varepsilon_2 a_{nj} + p_{nj}^{(t+1)}}{\varepsilon_2 + 1},$$

with

$$p_{nl}^{(t+1)} = \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}+1} p_{nl_{j_k}}^{(t+1)} \qquad p_{nj}^{(t+1)} = \sum_{l \in I,O} \sum_{k=1}^{K_{l_j}+1} p_{nl_{j_k}}^{(t+1)}.$$

Similarly to what is presented in Section 3.2.2, a Gaussian smoothing G_σ is applied to the responsibilities and the final update for the spatially varying mixing coefficients is

$$\begin{aligned}\tilde{b}_{nl}^{(t+1)} &= (1-\kappa_1) b_{nl} + \kappa_1 (G_\sigma \star p_{nl}^{(t+1)}) \\ \tilde{a}_{nj}^{(t+1)} &= (1-\kappa_2) a_{nj} + \kappa_2 (G_\sigma \star p_{nj}^{(t+1)}),\end{aligned}$$

where $\kappa_i = \dfrac{1}{\varepsilon_i + 1}$ and $\star$ represents the convolution operator.

4.2.2.2 Complete model

As detailed in Section 3.2.3, a spatial regularisation was added through a Markov Random Field at Level 2 and 3. The expression of the MRF with the matrix H allows for a different treatment between a Level 2 neighbourhood and a Level 3 neighbourhood. The constraint over the Gaussian covariances detailed in Section 3.2.4 was introduced in order to account for the difference between the biological variability of the tissues modelled through the various split operations and the noise due to the acquisition of the image itself and approximated as Gaussian throughout the full image. The bias field correction detailed in Section 3.2.1 was also applied to the log-transformed intensities. Within this framework, the E-step of the EM contributes to the update of the responsibilities so that

$$p_{nl_{j_k}}^{(t+1)} = \frac{\phi_{nl_{j_k}}^{\mathrm{c}(t)} \tilde{b}_{nl}^{(t)} \tilde{a}_{n_j}^{(t)} w_{l_{j_k}}^{(t)} \psi_{nl_{j_k}}^{(t)}}{\sum_{l' \in I,O} \sum_{j'=1}^{J} \sum_{k'=1}^{K_{j'}} \phi_{nl'_{j'_{k'}}}^{\mathrm{c}(t)} \tilde{b}_{nl'}^{(t)} \tilde{a}_{nl'_{j'}}^{(t)} w_{l'_{j'_{k'}}}^{(t)} \psi_{nl'_{j'_{k'}}}^{(t)}},$$

where the condensed notations

$$\begin{aligned}\phi_{nl_{j_k}}^{(t)} &= f\left(\mathbf{y}_n | \mathbf{z}_n = \mathbf{e}_{l_{j_k}}, \boldsymbol{\Xi}_{\mathbf{K}^{(t)}}\right) \\ \psi_{nl_{j_k}}^{(t)} &= \exp\left(-U_{\mathrm{MRF}}\left(\mathbf{e}_{l_{j_k}} | \mathbf{P}_{\mathbf{N}_n}^{(t)}, H\right)\right)\end{aligned}$$

are adopted. The remaining parameters can be updated as

$$w_{l_{j_k}}^{(t)} = \frac{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)}}{\sum_{k'=1}^{K_{l_j}+1} \sum_{n=1}^{N} p_{nl_{j_{k'}}}^{(t)}} \qquad \boldsymbol{\mu}_{l_{j_k}}^{(t)} = \frac{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} \mathbf{y}_n^{c^{(t)}}}{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)}}$$

$$\Delta_{l_{j_k}}^{(t)} = \frac{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} \left(\mathbf{y}_n^{c^{(t)}} - \boldsymbol{\mu}_{l_{j_k}}^{(t)}\right) \left(\mathbf{y}_n^{c^{(t)}} - \boldsymbol{\mu}_{l_{j_k}}^{(t)}\right)^T}{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)}}$$

$$\Psi^{-1(t)} = \frac{\sum_{l \in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}} S_{l_{j_k}}^{(t)} \Delta_{l_{j_k}}^{-1(t)} S_{l_{j_k}}^{(t)}}{N \cdot \sum_{l \in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}} 1} \qquad \Lambda_{l_{j_k}}^{(t)} = \frac{\Delta_{l_{j_k}}^{(t)} + \frac{S_{l_{j_k}}^{(t)} \Psi^{(t)} S_{l_{j_k}}^{(t)}}{\sum_n p_{nl_{j_k}}^{(t)}}}{1 + \frac{N+d+1}{\sum_n p_{nl_{j_k}}^{(t)}}}.$$

4.3 Algorithm implementation

The algorithm can be separated into three main steps:

- Preprocessing
- Initialisation
- Model selection

4.3.1 Preprocessing

The preprocessing of the data used in this work combines the most common steps evoked in Section 2.1. It consists first, when needed, in the spatial coregistration of the different modalities. When available, the T1 image is chosen as reference image due to the good contrast between healthy tissues observed in this modality. An affine transformation is used to register the modalities in order to avoid inconsistencies known to arise when applying non rigid registration to pathological data [69].

As detailed in Section 4.2.2.1, statistical atlases are used at Level 1 and 2 of the hierarchy. For Level 1, no *a priori* statistical information is known about the location of the outliers. The outliers are thus initially probabilistically defined according to the Mahalanobis distance as derived in Section 4.1.1.1 [195]. Therefore, the mixing proportions at Level 1 are initialised so that $\forall n, b_{nO} = b_O$ and $b_{nI} = 1 - b_O$. The value for b_O has been set to 0.01 in all our experiments, (equivalent to the probability of encountering a sample with Mahalanobis distance superior or equal to 3 when using two

modalities). Note however that this value does not correspond to the density estimate of a sample at Mahalanobis distance 3 from the mean of the distribution. As underlined by Van Leemput et al. [79], such a value would change according to each tissue distribution. Conversely, for Level 2, tissue-specific probabilistic maps are used to describe the prior probability of the four main tissues (GM, WM, CSF and NB). Two strategies can be adopted to obtain these probability maps. The first consists in choosing the smooth maps from the ICBM template (ICBM452) that are then aligned with the observed data and re-normalised between 0 and 1 before use. The alignment of the atlases is obtained by registration of a T1 atlas template to the target image using first an affine transformation [197] followed by a non-rigid registration [197]. The smoothness of the template helps preventing the problems mentioned earlier when using non-rigid registration on pathological data. The second strategy is to use the output of the label fusion of tissue segmentations obtained via application of the Geodesic Information Flow (GIF) pipeline [198].

Besides, for computational, normalisation and complexity purposes, images are roughly skull-stripped before segmentation. Here again two strategies have been used. The first one consists in the application of morphological operation on the brain extraction obtained via the method detailed in [199]. The masks obtained for brain extraction are filled to include the ventricles and sulcal CSF. This operation considers the CSF as part of the features to segment and may allow the segmentation of subarachnoid vessels. Furthermore, it has been noted that in the presence of large and highly pathological lesions, appearing as very hypointense on the T1 modality, errors in the brain extraction process can occur at the ventricles' borders. The initial elimination of some periventricular lesions can then be corrected by the morphological operations of dilation and filling. As a trade-off however, the dilation of the initial mask has been seen in some cases to contribute to the inclusion of some fat, skin, dura matter and bony structures neighbouring the external CSF, justifying further the inclusion of the NB class in the model. The second strategy is derived as for the atlas from the application of the GIF pipeline and is simply the output of the total intracranial volume (TIV) mask.The main contribution of the mask regards the process of image normalisation since the histograms on which the GMM will be fitted is defined at this stage. The potential presence of extreme intensity outliers outside of the mask region would greatly

increase the complexity of the segmentation if a mask was not used. It may indeed affect the normalisation and the shape of the histogram hindering the stability of the model evolution. The evaluation of the impact of these preprocessing choices is further detailed in Section 5.4.

4.3.2 Initialisation

The initial model $\mathbf{K}^{\text{init}}$ is initialised as $K_{I_j} = 1$ and $K_{O_j} = 0 \ \forall j$, meaning that each inlier tissue component is being modelled by a single Gaussian, and the outlier tissue components are modelled by a uniform distribution. As the inlier mixtures are assumed to be governed only by Gaussians, there is no uniform distribution for the inlier classes (weight fixed at 0). Thus, the initial model attributes one Gaussian component for each inlier tissue component and one uniform component for each outlier tissue component, allowing for a different treatment of inliers and outliers despite the flat priors. When building the mixtures for the outliers, the differences between inliers and outliers is driven by the differences in the statistical priors adapted after the first EM convergence.

To avoid overfitting, the bias field correction is only optimised on the initial model $\mathbf{K}^{\text{init}}$. Outliers that are not correctly detected during this first EM, may therefore contribute slightly adversely to the bias field modelling. Furthermore, to allow for a smoother and progressive modelling of the bias field, the maximal polynomial order of the basis functions is progressively increased. As far as the atlas adaptation is concerned, the adopted solution consists of relaxing the atlases only after convergence of the initial EM and considering them static afterwards. Practically, the segmentation obtained after convergence of the initial EM is smoothed and serves as atlas in the following steps of the process, as constantly evolving atlases within the model selection process could actually cause the instability in the model. Other strategies such as update and relaxation within the log-likelihood optimisation instead can also be preferred as well as alternative choices for κ. A constant value of $\sigma = 1$ has been used for all smoothing operations. The convergence threshold for the EM has been set at 10^{-6}.

4.3.3 Model selection

The model selection consists in the following steps:

Step 1 Computation of the list of possible SM model changes List_{SM}: given the current model, an ordered list of possible operations is defined by an alternating

sequence of splits and merges. Merges are ordered by increasing KLD and splits by decreasing KLD as detailed in Section 3.3.2.2. Priority is given to splits over uniform distributions. As a hard constraint, and mostly for computational reasons, Gaussian components with $w_{l_{j_k}} / \sum_{k'=1}^{K_{l_j}+1} w_{l_{j_{k'}}} < 0.01$ are not allowed to split. Merges can only occur between Gaussian components from the same mixture l_j.

Step 2 Initialisation of a new model: the first element of List_{SM} is used to initialise the current model according to Sections 3.3.2.1 and 4.1.2. The shape of the matrix H, used to define the MRF neighbourhood rules, is adapted to the new model.

Step 3 Optimisation of the tested model using the EM algorithm.

Step 4 Test of the new model using the BIC: the new model is accepted only if the relative change in the objective function is above 10^{-4}. For computational and stability reasons, Gaussian components with a relative weight $w_{j_{k_l}}$ below 0.01 are removed from the model.

Step 5 Conditional check: if the model is accepted, the process restarts at Step 1. If not, the first element of List_{SM} is removed from the list and the process restarts at Step 2, thus testing the next model change. If all the elements in the list have been tested, the algorithm terminates.

4.3.3.1 Remarks on the application of the BIC criterion

Practically, since the bias field is not updated after the first EM convergence, the BIC criterion developed in Section 3.3.2.3 is applied for the model selection without considering the bias field parameters as additional degrees of freedom. Alternative considerations over the number of degrees of freedom of the parameter space could account for:

- The strength of the priors over the covariance matrices.
- The weight of each subcomponent.
- The exchangeability of the Gaussian components for each tissue class with the inclusion of $\sum_{l \in I,O} \sum_{j=1}^{J} \ln(K_{l_j}!)$ in the number of degrees of freedom as mentioned by Bishop [200].

The model selection algorithm is graphically displayed in Figure 4.4.

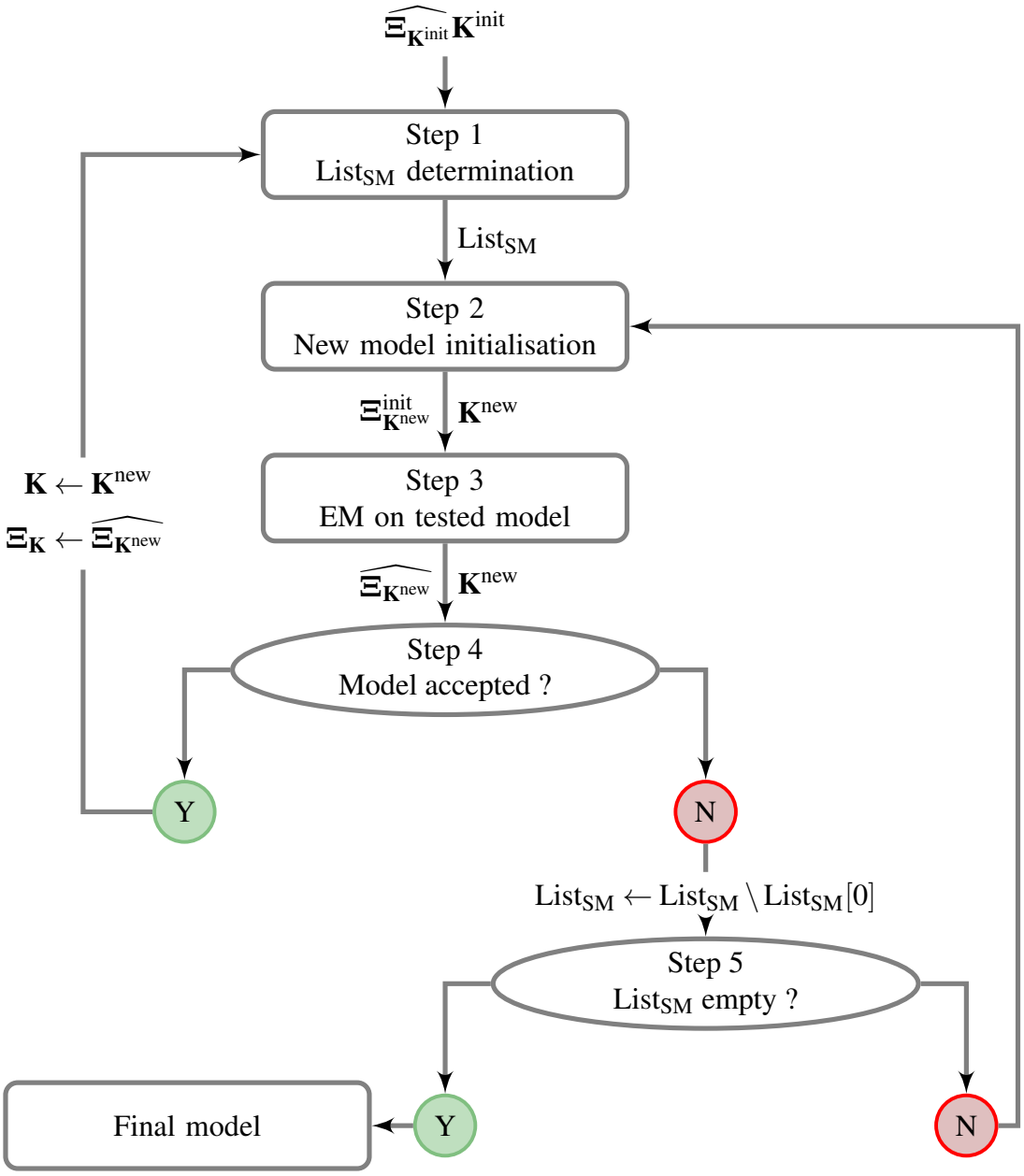


Figure 4.4: Graphic scheme of the model selection process performed in BaMoS.

Chapter 5

Application of BaMoS to white matter lesion segmentation and validation

In this chapter, the proposed data model is applied to segment white matter hyperintensities. The first part highlights the segmentation process and the correction for potential false positives. The remaining of the chapter is devoted to the validation of the WMH segmentation. Firstly, in Section 5.2, the various possible validation assessment measures are described and discussed. Then, after describing the different datasets in Section 5.3, the BaMoS lesion segmentations are internally evaluated in terms of robustness and performance in Section 5.4. Comparison to other existing algorithms is presented in Section 5.5.

5.1 WMH segmentation using BaMoS

5.1.1 Segmentation process

Once the appropriate model for the data has been selected, the result can be used to localise and delineate specific types of outliers such as white matter lesions. However, due to the high variability both in appearance and location of intensity outliers, WM hyperintensities can be modelled by more than one Gaussian component (Level 3 of the model). The selection of the appropriate clusters related to lesions can be obtained automatically based on intensity comparisons in a fashion similar to [79, 97, 101] as mentioned in Section 1.3.3. Eventually, those clusters are combined to produce the final lesion segmentation.

The model components relevant for lesion segmentation are selected from the out-

lier part of the model if spatially associated to the GM or the WM. Discarding the outlier part of the model associated with CSF or NB tissues prevents the inclusion of flow artefacts present in the CSF and most of the structures related to skin muscle or fat remaining in the mask. However, due to the smoothness of the GM and WM statistical atlases both tissue origins have to be further considered according to heuristic intensity comparisons rules stated hereafter:

$$l_{j_k} \in L \text{ if } \begin{cases} l_j = O_{\text{WM}} & \\ \text{or} & \text{and} \quad \mu_{l_{j_k}}^{(\text{FLAIR})} > \mu_{I_{\text{WM}}}^{(\text{FLAIR})} \\ l_j = O_{\text{GM}} & \mu_{l_{j_k}}^{(\text{T2})} > \mu_{I_{\text{WM}}}^{(\text{T2})} \\ & \mu_{l_{j_k}}^{(\text{PD})} > \mu_{I_{\text{WM}}}^{(\text{PD})} \end{cases},$$

with $Patho = \{\text{FLAIR, T2, PD}\}$ and L the set of possible lesion-related components.

When the FLAIR modality is not available, a further refinement on the selected components is required in order to avoid the inclusion of voxels related to partial volume effect at the GM-CSF border. Such components present themselves as very hypointense on the T1 modality and slightly but not strikingly hyperintense on the T2 image. Since the change in intensities is monotonous with lesion severity for both T1 and T2 intensities (respectively decreasing and increasing) but with a much stronger slope for the T2 compared to the T1 [51, 52], it can be assumed that very low intensities on the T1, would only correspond to a very severe lesion and thus an even higher hyperintensity level on the T2 image. To address this correspondence, in cases where a potential lesion-related component with a very hypointense mean on T1 (*i.e.* a mean lower that the mean observed for the GM inliers), the corresponding mean on the T2 was checked to be hyperintense compared to the mean of unsuspicious lesion components. Such components must present a hyperintense mean on T2 (lesion-like) and a mean on T1 higher than the mean of the GM inliers.

Mathematically, the set L of components that are presumed to be related to lesions is split into two groups: TL, standing for true lesion (or undisputed lesion), and DL, standing for disputed lesion such that

$$DL = \left\{ s \in L \;\middle|\; \mu_s^{(\text{T1})} < \mu_{I_{GM}}^{(\text{T1})} \right\}$$
$$TL = L \setminus DL.$$

The refined set of lesion RL is then defined as

$$RL = TL + \left\{ s \in DL \;\middle|\; \mu_s^{(T2)} \geq \mu_{TL}^{(T2)} \right\}.$$

where $\boldsymbol{\mu}_{TL}$ denotes the mean intensity of the TL set.

It must be emphasised that this refinement is not needed in the case where the FLAIR modality is available. Indeed, at the CSF-GM interface, intensities do not appear hyper-intense on the FLAIR image. Furthermore, the monotonous evolution of intensities with lesion severity does not hold in the case of the FLAIR modality due to the inversion recovery process of the acquisition [52]. A similar selection process is performed voxelwise for the remaining uniform distributions. The main strength of such selection process relies on its post-processing characteristic. As it occurs after obtaining the final model, it remains independent of the definition of the outliers of interest. Furthermore, this postprocessing step remains flexible and adaptable to differences in definitions according to the available modalities or clinical subtleties in fields that lack standard definitions [4]. It is for instance possible to include the CSF outliers and correct afterwards for the inclusion of false positives.

Figure 5.1 illustrates how BaMoS was able to separate different types of lesion in clinical data. Other types of outliers, such as areas of iron deposition in the basal ganglia with much darker intensities on FLAIR images were also assigned their own cluster. A visual example of BaMoS' ability to stratify different types of WM and deep GM sub-clusters is presented in Figure 5.1.

5.1.2 Weighting of lesion-related components

With the previous step, Gaussian components are selected to be consequently considered as related to the final lesion segmentation. However, naively combining these selected components cannot account for the fact that they have been split based on their intensities and therefore model different levels of hyperintensity or partial volume effect with the surrounding normal tissues in the image. To convey this refined information in the final segmentation result, a multiplicative downweighting of the Gaussian components has been introduced. Focusing on the modalities that convey information about the pathology (T2, FLAIR, PD), the lesion-related components whose Mahalanobis distance towards the mean of the WM inliers above 3 are weighted with the

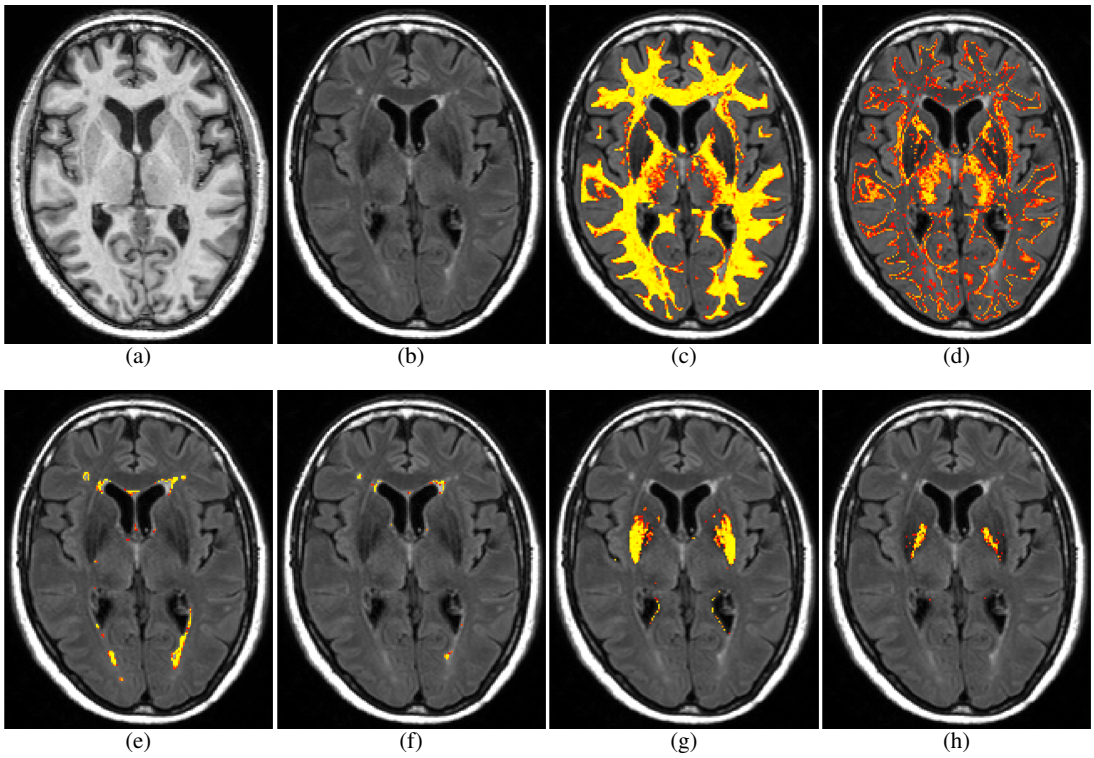


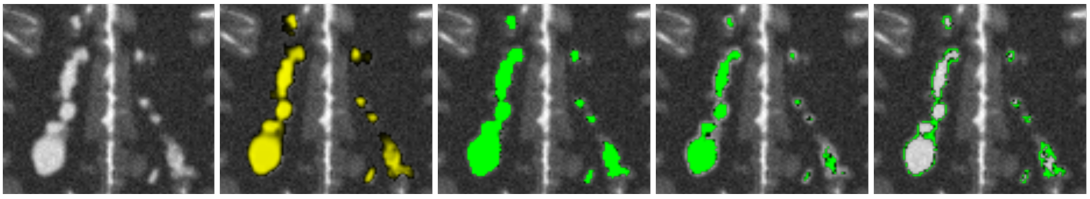
Figure 5.1: WM segmentation for clinical data. First row: Images of two modalities used T1 (a) and FLAIR (b) and the two subclasses obtained for the inliers of the WM (c-d). Second row: 4 subclasses classified under the WM part of the outliers with hyperintensities (e-f) and hypointensities (g-h).

maximum value 1. The others are attributed a lower weight according to the ratio of their Mahalanobis distance towards the WM inliers with respect to the threshold value 3 chosen for the weighting. Alternatively, the weighting can also be performed in a voxelwise fashion to avoid being dependent on the model structure. This is also the strategy adopted for the weighting in the uniform components. While this weighting accounts for lesion severity and partial volume, it should however be further refined in order to truly model the partial volume effect using for instance the contrast observed between WM and GM.

The lesion-related sub-components from the BaMoS model obtained from simulated data are displayed as an enlarged axial section in Figure 5.2. Note that different clusters are formed according to the intensity of the underlying voxel.

5.1.3 Correction for false positives

The tissue separation between outliers enables the correction for outliers that mimic WMH in the CSF due mostly to flow artefacts, but other regions were observed to be



(a) (b) (c) (d) (e)

Figure 5.2: Enlarged section on an axial slice of the T2-weighted (a) simulated image with severe lesion load. Overlayed with the lesion segmentation ground truth (b), the total segmentation obtained with BaMoS (c) and the two separated components lesion-related (d-e). Note that the separation between the lesion-related components is linked to the outlierness of the lesion.

prone to false positives (FP) after the simple lesion extraction step detailed in Section 5.1.1. Their occurrence was mostly observed in known controversial areas with ambiguous WMH definition [59] often described as normal occurrence in FLAIR images [57]. The septum pellucidum and its extension toward the genu and the splenium of the corpus callosum as well as the lining of the third ventricle were naturally classified as lesion-related whereas the manual segmentation omitted these regions. Due to the inclusion of the GM related outliers, FLAIR hyperintense areas in the cortical sheet were sometimes considered as lesions. Some of the aforementioned false positives can be easily corrected for. The correction for these misclassifications is based on three types of analysis:

Segmentation Refers to the binarisation of some classification in order to assess the lesions individually.

Lesion characterisation Refers to the features associated with each lesion used for their classification.

Lesion classification Refers to the final taxonomy associated to each lesion according to the features previously derived.

5.1.3.1 Segmentation previous to false positives correction

The following steps are performed:

Global segmentation The global segmentation for the anatomical tissues under study $T = \{\text{GM, WM, CSF, NB}\}$ is obtained by first combining the subclasses of each general tissue class. As far as the outlier components are concerned, they are naturally associated with their inlier counterparts. The lesion segmentation obtained

previously is reassigned to the WM. In the case of pronounced atrophy, it may also happen, that part of the CSF is more spatially associated with WM for which the corresponding voxels are outliers. The associated tissue for these elements is easily corrected to obtain a more appropriate tissue segmentation. For each of the tissues in T, the corresponding classification based on maximum probability is denoted S_T. This roughly corrected segmentation remains however suboptimal since some subclasses that may arise due to partial volume between tissues are binarily associated to one or the other tissue, only based on their Level 2 belonging. Furthermore, the corrections performed here in a post-processing manner may modify the model obtained after convergence.

Lesion binary segmentation The binary lesion segmentation can also be performed in different ways that are listed below from the more conservative to the least conservative:

Thresholding $z_n = \text{lesion}$ if $p(z_n = \text{lesion}) > 0.5$

Tissue segmentation $z_n = \text{lesion}$ if $p(z_n = \text{lesion}) > \max_T p(z_n = T)$ with $T \in \{\text{GM}, \text{WMI}, \text{CSF}, \text{NB}\}$, WMI being the white matter inlier tissue.

Subpart of WM segmentation $z_n = \text{lesion}$ if $S_{\text{WM}}(n) = 1$ and $p(z_n = \text{lesion}) > \max_{k \in [1; K_{I_{\text{WM}}}]} p(z_n = I_{\text{WM}_k})$

The differences in volumes seem very marginal but a more careful assessment is still needed.

Ventricles segmentation Using the statistical atlas corresponding to the internal CSF, a rough segmentation is obtained applying the following rule:

$$z_n = \text{Ventricle if } S_{\text{CSF}}(n) \times A_{\text{ICSF}}(n) > 0$$

where A_{ICSF} denotes the statistical atlas corresponding to the internal CSF.

5.1.3.2 Lesion characterisation for false positives correction

Once the overall binarised lesion segmentation is obtained, the connected components (CC) are determined applying the algorithm detailed by Borgefors et al. [201]. For each

of the connected components (*i.e.* each of the individual lesions), various measures related to its location and the surrounding tissue segmentation is assessed.

As far as the neighbourhood characteristics are concerned, they can be classified in different subclasses :

Distance

- Distance for each direction between the centre of gravity of the lesion and the centre of gravity of the mask that is the $\left\{\mathrm{CG}_{\mathrm{Lesion}}(t) - \mathrm{CG}_{\mathrm{Image}}(t)\right\}$, where CG denotes the centre of gravity and t one of the three Euclidean axis.
- Closest distance to the ventricle based on the ventricle segmentation performed in Section 5.1.3.1 that is $\min_{b_l \in B_l, b_v \in B_v} d(b_l, b_v)$ where B_l (resp. B_v) denotes the border of the lesion (resp. the border of the ventricles).
- Distance to each of the tissues based on the classification of the inlier part of the model only.

Neighbourhood

- Binary statement of neighbourhood between tissues
- Proportion of voxels on the external lesion border that belong to one of the 4 inlier tissues or to the outliers

Specific location Based on the initial statistical atlases of deep gray matter (DGM) A_{DGM} and internal CSF (ICSF) A_{ICSF}

- Possibility to be in the DGM if $\exists n \in \mathrm{CC}_l | A_{\mathrm{DGM}}(n) > 0$
- Proportion of voxels in the lesions with high probability (>0.45) to be inside the ventricles

5.1.3.3 Potential lesions taxonomy - False positives correction

Based on the descriptors defined in Section 5.1.3.2, each of the individual connected components was finally classified in a taxonomy of regions accepted or not as lesions. The taxonomy is:

Lesions

Periventricular lesion: Lesion neighbourhing the ventricles, with most of their neighbours belonging to the normal appearing white matter.

Subcortical lesion: Lesion completely embedded in white matter or located in the area of the deep gray matter, but not in the medial part of the brain.

Subcortical lesion close to cortical sheet: Lesion with neighbours belonging to the GM but with a higher proportion of WM neighbours and not belonging to the DGM.

Discarded regions

Out region: All regions of hyper-intensities close to the image mask. It concerns mostly the voxels related to muscle, fat or skin remaining in the mask.

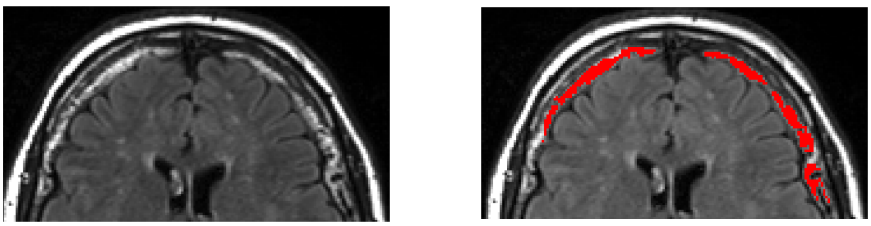


Figure 5.3: Example of the out region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

Septum Pellucidum In those regions, the proportion of neighbours belonging to the three main tissues is quite high and the distances toward the midline and the centre of gravity of the mask relatively low.

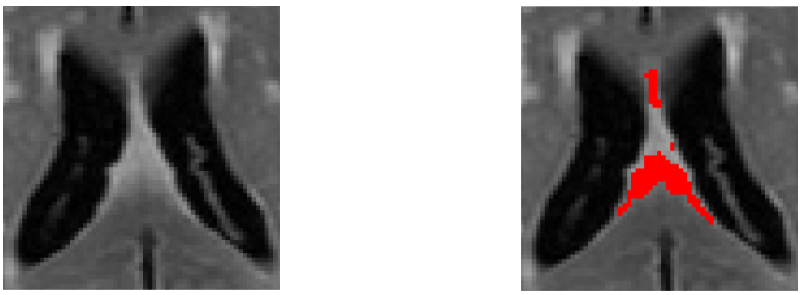


Figure 5.4: Example of the septum pellucidum/corpus callosum region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

Cortical Sheet Regions with none or almost no neighbours in the white matter, not close to the ventricles and not belonging to the susceptible deep gray matter area.

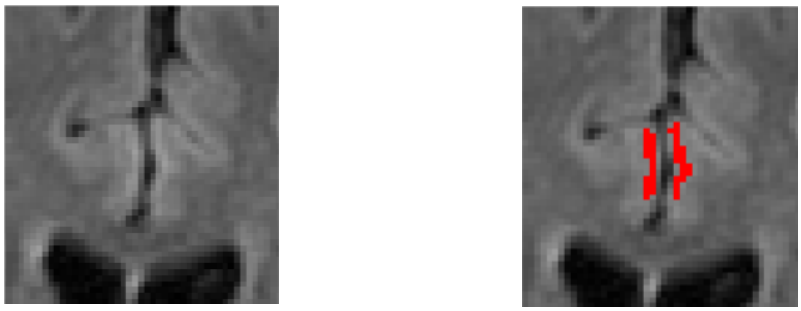


Figure 5.5: Example of the cortical sheet region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

Third ventricle Areas close to GM, WM and CSF for which the cortical sheet classification does not apply the proportion of WM neighbours is low compared to GM, and the location of the centre of gravity close to the midline is more temporal than the centre of gravity.

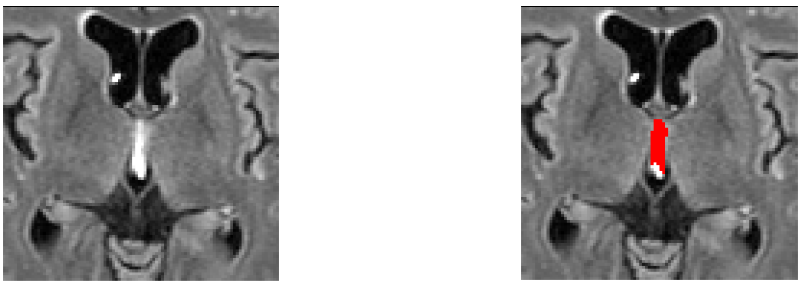


Figure 5.6: Example of the third ventricle region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

Fourth ventricle Regions for which the proportion of neighbours belonging to WM is lower than those of GM, whose distance to the CSF is lower than the distance to the ventricles and whose centre of gravity is lower than the centre of gravity of the mask.

Figure 5.7: Example of the fourth ventricle/Sylvean aqueduct region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

Choroid plexus/Inside Ventricle Regions for which the proportion of WM neighbours is very low, the location is compatible with being in the ven-

tricles and the proportion of surrounding CSF voxels is higher than the proportion of GM surrounding voxels.

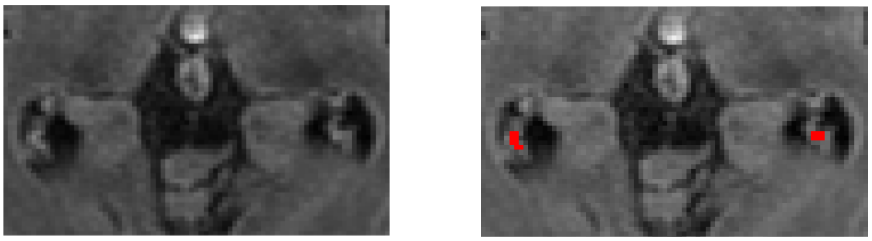


Figure 5.8: Example of the choroid plexus region correction, with the FLAIR image on the left and the FP to be corrected (red) on the right.

5.2 Lesion segmentation assessment: measures and limitations

Validation of automated segmentation methods is usually performed based on a gold standard (GS) that is difficult to acquire for clinical data and is commonly based on manual segmentation. As the total volume load is currently the clinical standard for the assessment of lesions, correlations between the automated and manual lesion volumes, here considered as a clinical gold standard, is a common form of validating and comparing segmentation strategies [79, 90, 108]. Contrarily to the lesion count, lesion volume has been found to be related to the clinical outcome and to cognitive decline [62] when investigating age-related white matter changes

5.2.1 Volumetric assessments

Since the volume is a popular way of assessing the lesion burden, the Pearson's correlation coefficient between volumes has been widely used to compare lesion segmentation performance on a population. An other evaluation measure directly related to volume, that is applied on an individual case basis is the volume difference (VD) expressed in percent and defined as $100\left|1-\frac{\sharp_v \mathrm{Seg}}{\sharp_v \mathrm{Ref}}\right|$, where Seg refers to the lesion segmentation, Ref to the gold standard and $\sharp_v$ to the number of voxels in the considered set. Generally, $\sharp$ denotes the cardinality of a set. This measure suffers from the fact that it is not defined for a null reference volume, that is for a normal case in which no lesion has been segmented. However, as noticed by Schmidt et al. [90], such volume assessments are insufficient to assess the quality of the segmentation since a good agreement

in terms of volume does not mean that the segmentation in itself is accurate. A score of 0 for the volume difference can be obtained without any overlap between the two compared segmentations.

5.2.2 Dice Similarity Coefficient: Popular but limited

As a measure of spatial overlap, the Dice Similarity coefficient (DSC), defined as

$$\mathrm{DSC} = \frac{2\sharp_v(\mathrm{Ref} \cap \mathrm{Seg})}{\sharp_v \mathrm{Ref} + \sharp_v \mathrm{Seg}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

is a very popular measure of agreement. However its application to the lesion segmentation context has been challenged.

As evoked by Schmidt et al. [90] and Anbeek et al. [80], the DSC is dependent on the lesion load: the same volumetric amount of misclassification will affect more the small than the large lesion loads. In addition, the same volumetric amount of error, outline errors will affect differently the results according to error type, shape and number of lesions, modifiers of the surface/volume ratio and of the volume per lesion.

5.2.2.1 Error configurations - Impact of individual lesion volume

Simplistic examples allow for a better illustration of this effect and the impact on the similarity coefficient. Assuming isotropic images with unit voxel element, individual lesions are simple shapes characterised by a single length a. The volume of an individual lesion is denoted V_a. With a systematic error of one voxel at the outline of each individual lesion to segment, three error configurations are studied. These three examples of error configurations are described hereafter:

UnderSeg Each individual lesion of volume V_a is undersegmented with an outline error of one voxel thus resulting in a volume V_{a-2}. See Figure 5.9 b.

OverSeg Segmentation resulting from the systematic overestimation of the individual reference lesions by one voxel at the outline, resulting in individual lesions of shape characteristic $a+2$ of volume V_{a+2}. See Figure 5.9 c.

TransSeg Segmentation resulting from the translation along the main diagonal by one voxel of the reference lesion producing individual lesions of volume V_a. See Figure 5.9 d.

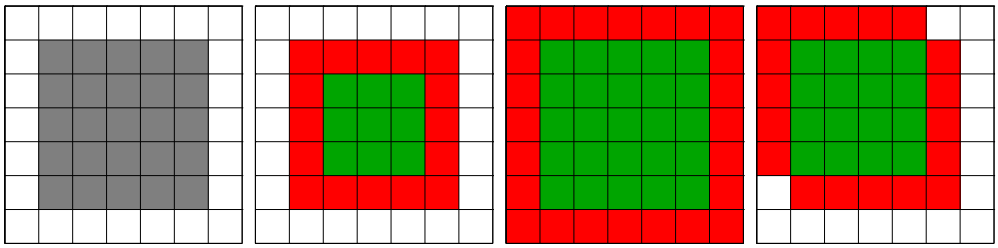


Figure 5.9: Presentation of the three studied configurations of errors UnderSeg (b), OverSeg (c) and TransSeg (d) with respect to the segmentation of reference (a). In the compared segmentations, green refers to TP, red to erroneous segmentation (FN or FP) and white to true negatives.

The three studied configurations are shown in Figure 5.9 and the type and amount of errors are recapitulated in Table 5.1.

The influence of the error on the DSC is then expressed for the three configurations as follows:

UnderSeg	OverSeg	TransSeg
$\text{FN} = V_a - V_{a-2}$	$\text{FP} = V_{a+2} - V_a$	$\text{FN + FP} = 2V_a - 2V_{a-1}$
$\text{TP} = V_{a-2}$	$\text{TP} = V_a$	$\text{TP} = V_{a-1}$
$\text{DSC} = \frac{2V_{a-2}}{V_a + V_{a-2}}$	$\text{DSC} = \frac{2V_a}{V_a + V_{a+2}}$	$\text{DSC} = \frac{V_{a-1}}{V_a}$

The DSC is therefore linked to the volume of an individual lesion, the DSC being much more affected by errors at the border for small individual lesions as illustrated in Figure 5.10.

5.2.2.2 Shape impact

It can also be noted that the shape chosen for individual lesions in this simplistic model affect the DSC to different extents. Considering now individual lesions of spherical shape, their individual number of elements when considering a radius a is $V_a = \frac{4\pi a^3}{3}$

Seg	Volume	TP	FP	FN
UnderSeg	V_{a-2}	V_{a-2}	0	$V_a - V_{a-2}$
OverSeg	V_{a+2}	V_a	$V_{a+2} - V_a$	0
TransSeg	V_a	V_{a-1}	$V_a - V_{a-1}$	$V_a - V_{a-1}$

Table 5.1: Errors and volumes for an individual lesion of cubic shape with three configurations of outline error.

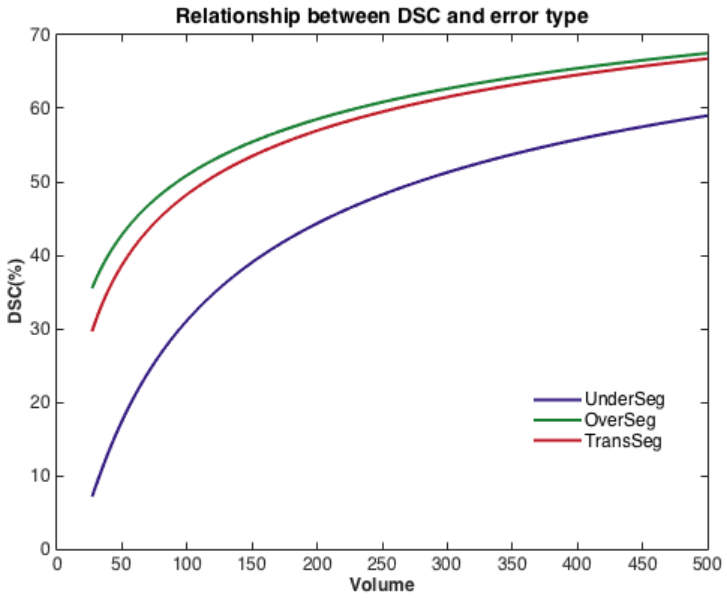


Figure 5.10: Influence of the individual volume and the type of segmentation error on the DSC in a cubic configuration.

and the corresponding surface is $S_a = 4\pi a^2$. Similarly to the previous derivation, with an error at the border of one element as in the OverSeg configuration, the DSC becomes in the spherical case:

$$\begin{aligned} \mathrm{DSC}_{\mathrm{spheric}} &= \frac{2\mathrm{V}_a}{\mathrm{V}_a + \mathrm{V}_{a+1}} \\ &= \frac{2 \cdot \mathrm{V}_a}{V_a + \frac{4\pi}{3}\left(\left(\frac{3\mathrm{V}_a}{4\pi}\right)^{1/3} + 1\right)^3}. \end{aligned}$$

When considering a regular tetrahedron with a side of length a, and of volume $V_a = \frac{\sqrt{2}}{12}a^3$ the DSC can in turn for the same type of outline error be expressed as

$$\begin{aligned} \mathrm{DSC}_{\mathrm{tetrahedric}} &= \frac{2\mathrm{V}_a}{\mathrm{V}_a + \mathrm{V}_{a+2}} \\ &= \frac{2\mathrm{V}_a}{\mathrm{V}_a + \frac{\sqrt{2}}{12}\left(\left(\frac{12\mathrm{V}_a}{\sqrt{2}}\right)^{(1/3)} + 2\right)^3}. \end{aligned}$$

Again, the DSC is directly linked to the volume of individual lesion but the equation differs due to the shape of the individual lesions considered. Table 5.2 summarises the effect in Dice score for different lesion size (in number of elements) for the cu-

Number of voxels per individual lesion	DSC (%) Cubic	Spheric	Tetrahedric
27	35.5263	43.1657	59.9524
125	53.4188	60.3769	73.7665
1000	73.3138	77.9502	86.0601
3375	81.4431	84.8073	90.5300
8000	85.8001	88.4247	92.8327
27000	90.3494	92.1652	95.1794
125000	94.1237	95.2439	97.0876

Table 5.2: Effect on the DSC (%) of a one element border error in overestimation according to shape and volume of individual lesions.

bic, spherical case and tetrahedrical case while Figure 5.11 displays graphically the corresponding results.

5.2.3 Voxelwise and cardinal assessment

Since global volumetric comparison may be limited in their assessment of the actual overlap with a reference segmentation and the use of a unique overlap assessment measure may be biased by the ground truth characteristics, other ways of assessing the quality of a given segmentation have been developed. As underlined in both reviews by Lladó et al. [50] and García-Lorenzo et al. [110], there is no unique scalar evaluation measure able to summarise globally the quality of the segmentation and these measures

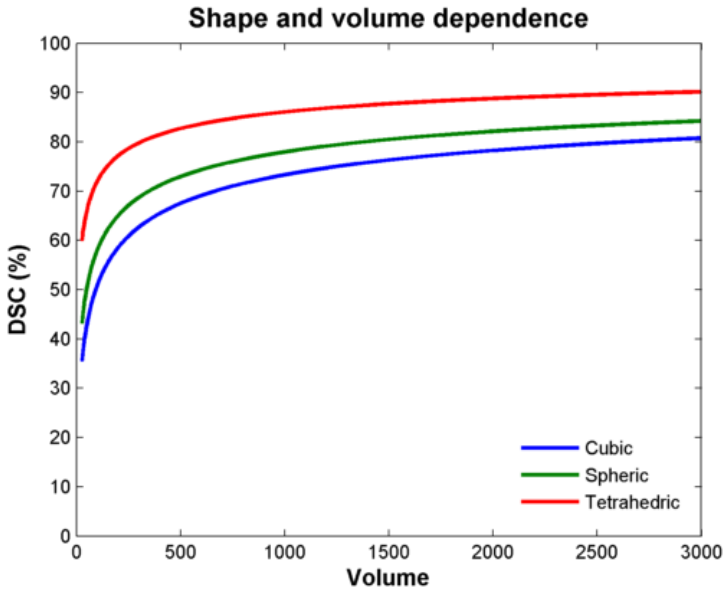


Figure 5.11: Dependence of the DSC on shape and volume of lesion when considering a systematic overestimation by one voxel at the border (OverSeg).

have strengths and limitations. Their combined use may not only help in the evaluation of a given segmentation with respect to a reference but also allow to better understand the origin of potential disagreements. Such strategy has been used for the MICCAI 2008 challenge on MS segmentation [159].

Additionally, these measures may be apply either at the voxel level or at the cardinal level. When needed the subscripts v and c are used to distinguish between these two levels. At the cardinal level, a lesion or connected lesion component is defined as a maximal set of spatially connected lesion voxels. The degree of connectivity is let to the user discretion. True positives are then lesions that share at least one voxel in the Seg and Ref images. A false negative is a lesion whose voxels are only classified as lesion in the Ref image. On the contrary, none of the voxels of a false positive lesion can be found in the Ref image. The sets of true positive, false positive and false negative lesions are respectively denoted TP_c, FP_c and FN_c.

Three families of measures can then be separated to assess the segmentation quality:

- Purely voxelwise,
- With voxel and cardinal definition,
- Mixing cardinal and voxelwise definitions.

Purely voxelwise

Average distance (AvDist) The average distance, mentioned by Datta et al. [99] and Styner et al. [159] measures the average distance between the two lesion outlines

$$\mathbf{AvDist}(\mathrm{Ref},\mathrm{Seg}) = \frac{\sum_{s\in\partial\mathrm{Seg}} \min_{s\in\partial\mathrm{Seg}} d(s,r) + \sum_{r\in\partial\mathrm{Ref}} \min_{s\in\partial\mathrm{Seg}} d(s,r)}{\sharp_v\partial\mathrm{Seg} + \sharp_v\partial\mathrm{Ref}}$$

where $\partial\mathrm{Seg}$ (resp. ∂Ref) denotes the border in the 18-neighbour connectivity of the Seg (resp. Ref) set. and $d(s,r)$ is the Euclidean distance between element s and r. It must be however noted the difficult definition of such an assessment measure when one of the volumes is 0.

With voxel and cardinal definition

True positive rate (TPR) The true positive rate (TPR) that can be defined either at the voxel or cardinal level is expressed as $\frac{\sharp \mathrm{TP}}{\sharp \mathrm{Ref}}$ and takes its values in [0 ; 1] with 1 as best value. With this measure, a perfect score at the voxel level can be reached for a suboptimal segmentation if the errors are exclusively false positives. In the cardinal form, the ratio becomes dependent of the lesion spatial connectivity since joined lesion are only counted once.

False positive rate (FPR) The false positive rate, as the TPR expressed as a ration taking its values in [0 ; 1] (best value 0), can also be defined in its voxelwise or cardinal version and is expressed as $\frac{\sharp \mathrm{FP}}{\sharp \mathrm{Seg}}$. Both the voxelwise and the cardinal definition reach their limit when no false lesion is detected. However, empty results are not defined for this measure.

False negative rate (FNR) The false negative rate, also taking its values in [0 ; 1] (best value 0) is again to be defined voxelwise or cardinalwise and is expressed as $\frac{\sharp \mathrm{FN}}{\sharp \mathrm{Ref}}$ or $1 - \mathrm{TPR}$.

It must be additionally noticed that the cardinal version of TPR and FPR, that is using the number of connected lesions instead of the number of voxels, gives a similar weight to the connected lesions independently of their volume. Thus errors made on very small lesions, that are generally harder to detect will be emphasised, whereas massive errors in the outline of bigger lesions will be overseen. Similarly to the volume difference problem, it must be noted that neither TPR nor FPR are defined when the amount of ground truth lesions ($\sharp$Ref) is 0. The DSC detailed earlier can also be classified under this category.

Mixing voxel and cardinal definitions Recently, new inter-rater assessment measures with application to MS lesion segmentation have been developed. These have shown to be less dependent than the DSC to the assessed lesion burden [202].

Detection error (DE) The detection error is the volume of error measured cardinally and is expressed as

$$\mathrm{DE} = \sum_{F \in \mathrm{FP}_c} \sharp_v \mathrm{Seg}_F + \sum_{F \in \mathrm{FN}_c} \sharp_v \mathrm{Ref}_F.$$

Name	Equation	Best	Range
DSC	$100 \times \frac{2\sharp(\text{Ref}\cap\text{Seg})}{\sharp\text{Ref}+\sharp\text{Seg}}$	100 (%)	[0 ; 100]
VD	$100 \times \left\lvert 1 - \frac{\sharp\text{Seg}}{\sharp\text{Ref}} \right\rvert$	0 (%)	[0 ; ∞]
FPR	$\frac{\sharp\text{Seg}-\sharp(\text{Ref}\cap\text{Seg})}{\sharp\text{Ref}}$	0(%)	[0 ; 100]
TPR	$100 \times \frac{\sharp(\text{Ref}\cap\text{Seg})}{\sharp\text{Ref}}$	100 (%)	[0 ; 100]
FNR	$\frac{\sharp\text{Ref}-\sharp(\text{Ref}\cap\text{Seg})}{\sharp\text{Ref}}$	0(%)	[0 ; 100]
AvDist	$\frac{\sum_{s\in\partial\text{Seg}} \min_{r\in\partial\text{Ref}} d(s,r) + \sum_{r\in\partial\text{Ref}} \min_{s\in\partial\text{Seg}} d(s,r)}{\sharp\partial\text{Ref}+\sharp\partial\text{Seg}}$	0 (mm)	[0 ; ∞]
DE	$\sum_{F\in\text{FP}_c} \sharp_v\text{Seg}_F + \sum_{F\in\text{FN}_c} \sharp_v\text{Ref}_F$	0 (mL)	[0 ; ∞]
OER	$100 \times \frac{\sum_{T\in\text{TP}_c} \sharp_v(\text{Ref}_T\cup\text{Seg}_T) - \sharp_v(\text{Ref}_T\cap\text{Seg}_T)}{\sharp_v\text{Ref}}$	0 (%)	[0 ; ∞]

Table 5.3: Table of lesion segmentation evaluation measures.

Outline error rate (OER) The OER is measured as the ratio between the volume of voxelwise error found for the true positive components and the Ref volume.

$$\text{OER} = \frac{\sum_{T\in\text{TP}_c} \sharp_v(\text{Seg}_T \cup \text{Ref}_T) - \sharp_v(\text{Seg}_T \cap \text{Ref}_T)}{\sharp_v\text{Ref}}.$$

Table 5.3 gathers the measures used for assessment in the BaMoS experiments in comparison to other algorithms along with the best possible value obtained for each of them. In order to better assess the origin of the errors, based on the definition by [202], additional evaluation may be carried out:

OE/TotF Measures the proportion of total error (TotF) that is related to the outline error.

FP/TotF Measures the proportion of error that is false positive.

OEFP/FP Measures among the false positives, the proportion that relates to the outline error.

OEFN/FN Measures among the false negatives, the proportion that relates to the outline error.

5.2.4 Consistency appraisement

Although they account for different aspects of the segmentation assessment, all the previously presented measures rely on the existence of a reference segmentation. Gold standard based assessments evaluate the algorithm's ability to reproduce human behaviour rather than its true ability to detect abnormal biological signal. One should thus be cautious when assessing an algorithm entirely based on manual segmentations since the high inter- and intrarater variability may lead to inconsistencies in the reference with respect for instance to the intensities. FLAIR intensities standardised with respect to the normal appearing WM (NAWM) also known as Z-scores can be used to assess such intensity consistency. The FLAIR Z-score with respect to the WM inliers for voxel n is defined as $Z = \frac{y_n^{\text{FLAIR}} - \mu_{\text{IWM}}^{\text{FLAIR}}}{\sigma_{\text{IWM}}^{\text{FLAIR}}}$. In order to check intensity consistency, the following evaluations can be introduced:

PropLes : Proportion of segmented lesion whose intensity Z-score range overlaps with the one of the NAWM.

PropWM : Proportion of WM whose intensity Z-score range overlap with the one of the segmented lesion

DistMin : Difference in Z-score between the minimal intensity segmented as lesion and the maximal intensity considered as normal WM.

DistQuant : Difference in Z-score between the first quartile of intensity Z-score for the lesion segmentation and the third quartile of the intensity Z-score for the WM segmentation.

5.3 Data description

In order to investigate the behaviour of BaMoS in a large set of conditions and experiments, different datasets were used and their main characteristics and strengths are detailed in this section.

5.3.1 Brainweb

García-Lorenzo et al. [110], consider the use of synthetic data for the validation of any lesion segmentation algorithm as an imperative step in any validation framework. Publicly available, the MR simulated Brainweb brain images `http://brainweb.`

bic.mni.mcgill.ca were used. This simulator allows for a model of MS lesions at three different lesion loads (Mild, Moderate, Severe). The ground truth segmentations are provided as maps of fuzzy membership. The simulator makes available 3 modalities (T1 T2 and PD) and allows for different level of noise and bias field intensity. The validation was performed on all combinations of the modalities, at various noise levels (3%, 5% and 7%) and at different strengths of intensity inhomogeneity (0%, 20% and 40%) for the three available lesion loads (0.4 mL, 3.5 mL, 10.1 mL). Since the quality of imaging data is very heterogenous, assessing the robustness of a given method against the level of noise is an important validation step. The range of noise between 3% and 7% (as defined in Brainweb) was found to be comparable to the range of noise of 3T and 1.5T clinical scans, and was thus used for comparison. The noise model in Brainweb consists in adding Gaussian white noise on both the real and imaginary components of the image with a standard deviation chosen based on a reference tissue signal such that the ratio between the standard deviation and the signal is the percentage value of the noise model. The obtained noise on the magnitude image thus follows a Rician distribution. As the signal for the reference tissue varies through modalities, the observed effect of noise is also modality-dependent. For the binary segmentation assessment, the ground truth maps were thresholded at 0.5. The main strength of such synthetic data is the existence of an indisputable ground truth which is not the case for any manual segmentation. It is complemented by the possibility to test for various image quality conditions (noise and bias field).

5.3.2 T2DB (Type 2 diabetes)

Type 2 diabetes has been reported as a risk factor for the presence of WMH in ageing populations [203]. For this study, brain images from Type 2 Diabetes patients and matched controls with increased cardiovascular risk (age $>$ 50) were acquired on a 3T Philips scanner. Multi-slice FLAIR images ($0.958 \times 0.958 \times 3$ mm^3) and T1-weighted 3D registered images were used. Further details about the acquisition and preprocessing can be found at http://mrbrains13.isi.uu.nl. WMH were manually segmented on twenty FLAIR images giving a total lesion load (TLL) range between 0 mL and 35.48 mL (median 6.02 mL, interquartile range 9.22 mL) and used as gold standard for the evaluation of the automated methods. No WMH was detected

by the human rater for one of the twenty subjects.

5.3.3 MICCAI MS

The MICCAI MS Challenge 2008 (`http://www.ia.unc.edu/MSseg/`) dataset was used to validate BaMoS for MS. For this dataset, 20 T1, T2 and FLAIR images are provided along with the corresponding manual segmentations. All images are re-sampled isotropically to the space of the T1 image. Comparison between methods was performed using T1 and FLAIR images. This dataset is the only one dedicated to MS and therefore allows for the study of an application apart from ageing population WMH. Additionally manual segmentation of the lesions are available allowing for a comparison across algorithms.

5.3.4 ADNI 92

Part of a starting study in the context of manual protocol segmentation evaluation, 92 datasets with T1 and FLAIR pulse sequence from the ADNI database were selected. The ADNI (Alzheimer's disease Neuroimaging Initiative) project was launched in 2003 as a public-private partnership, whose primary goal has been to test whether serial magnetic resonance imaging, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Information about the study can be found at `www.adni-info.org`. Being a multi-centre study, these 92 scans were acquired at eight different sites, spanning three MR manufacturers. In the ADNI protocol, the T1 images are of resolution 1.19 $\times$ 1.0 $\times$ 1.0 mm^3 and the FLAIR images are of resolution 0.89 $\times$ 0.89 mm^2 in the axial plane with 5 mm slice thickness. The main strengths of such a dataset are its variability in terms of acquisition conditions for a given protocol and the large slice thickness that allows for testing on low resolution images. Furthermore, volumes of WMH, segmented according to [204] are made available online allowing for a partial comparison.

5.3.5 POPPY

The data used came from 71 subjects scanned as part of the POPPY study that investigates the relationship between HIV status, cardiovascular risk factors and neuroimag-

ing findings. With a mean age of 58.5 years, this mixed population of HIV positive and negative (54% HIV+) underwent brain MRI. Among the sequences, T1-weighted and FLAIR images with both 2D Axial and 3D protocols were acquired. MR protocol acquisition parameters with the resolutions and field of view (FOV) given in the order Foot Head (FH), Anterior Posterior (AP) and Left Right (LR) in mm and the times in ms were the following:

T1-weighted TR (repetition time) = 6.6 ms; TE (time to echo) = 3.1 ms; voxel size = 1.11 $\times$ 1.11 $\times$ 1.12 mm^3; FOV = 270 $\times$253 $\times$ 203 mm^3

FLAIR 2D (FLAIR2) TR = 8000 ms; TI (time at inversion pulse) = 2400 ms; TE = 125 ms; voxel size: 3.0 $\times$ 1.0 $\times$ 1.0 mm^3; FOV = 240 $\times$ 180 $\times$ 150 mm^3

FLAIR 3D (FLAIR3) TR = 4800 ms; TI = 1650 ms; TE = 282 ms; voxel size = 1.04 $\times$ 1.04 $\times$ 0.56 mm^3; FOV = 250 $\times$ 267 $\times$ 180 mm^3

The most important strength of this dataset is the joint availability of 2D and 3D FLAIR images acquired during the same MR scanning session.

5.3.6 SABRE

This cohort study is based on a tri-ethnic population and aims to assess the risks of diabetes and cardiovascular disease including small vessel disease in the brain [205]. All participants underwent MRI according to a standard protocol on a Philips Achieva 3.0-Tesla scanner. The series included the following imaging sequences:

T1-weighted 3D sagittal T1-weighted FFE (Fast Field Echo): TR = 6.9 ms; TE = 3.1 ms; voxel size = 1.0 $\times$ 1.0 $\times$ 1.0 mm^3; FOV = 256 $\times$ 256 $\times$ 180 mm^3

T2-weighted 3D sagittal T2-weighted TSE (Turbo spin echo): TR = 2500 ms; TE = 222 ms; voxel size = 1.0 $\times$ 1.0 $\times$ 1.0 mm^3; FOV = 256 $\times$ 256 $\times$ 180 mm^3

FLAIR sagittal 3D FLAIR (Fluid attenuated inversion recovery): TR = 4800 ms; TI = 1650 ms; TE = 125 ms; voxel size = 1.0 $\times$ 1.0 $\times$ 1.0 mm^3; FOV = 256 $\times$ 256 $\times$ 180 mm^3

An imaging data comprising 84 participants was used with mean age 71.3 years (SD=5.7). All images were reviewed for incidental pathology and scan quality. Two participant's scans were discarded from the analysis due to severe motion artefacts.

5.4 Internal consistency validation

Generally an algorithm is built to answer specific questions or requirements and choices of parameters or fixed inputs may strongly alter the results. As mentioned in Chapter 2, parameters such as discriminative thresholds can be tuned by cross-validation based in the case of lesion segmentation on databases of manual segmentations. For instance, Wang et al. [124] use this strategy for the choice of the trimming parameter while Schmidt et al. [90] recommend it to decide on the initial threshold to apply on the belief maps. When a few scalar parameters are involved [89, 113, 120], the space of solutions is jointly optimised against databases of manual segmentations. Others have investigated the influence of the choice of a given scalar parameter, fixing all the others, on the robustness of their proposed solution. In such investigations as performed by Van Leemput et al. [79] or García-Lorenzo et al. [101], the parameter of interest is discretely sampled within a range and a new segmentation result is obtained for each of them. If those parameter choices can be of paramount importance in the performance of the given method, other aspects of the process appear also to affect greatly the outcome. In methods requiring the normalisation of the intensities for comparison purposes, Steenwijk et al. [118] present evidence that the initial choice of normalisation strategy affects the results. Skull-striping strategies, initialisation choices with statistical priors, registration methods and parameters are as many aspects, often indirectly related to the core of the method and blamed for failure cases, whose impact is rarely investigated despite its potential importance. Other considerations that may affect the results are the reference space/discretisation in which the images are segmented or the choice of pulse sequences whose influence is notably studied by Guizard et al. [112] and Dyrby et al. [81]. As underlined in Section 2.6, the question of robustness beyond the simplified question of noise level and bias field strength, as can be evaluated with the Brainweb database (cf Section 5.3.1), can be studied under the perspective of clinical trials. This implies the evaluation of the acquisition protocol influence as well as the image resolution. The impact of the resolution on the segmentation results is for instance evaluated by Schmidt et al. [90] using resliced images. Finally, postprocessing strategies may further affect the final results. Comparisons of postprocessing solutions are for instance performed by Steenwijk et al. [118] on the topic of the minimal lesion size and in Gibson et al. [156] on the probabilistic threshold for the WM mask and the

connectivity to the template.

In the case of the BaMoS framework, the list of potential sources of variability in the outcome is long and investigating the joint space of such variations and their interactions would prove intractable. Among the scalar parameters, choices of the outlier initial probability map or energy choices in the MRF matrix come quickly to mind. The choice for the convergence criteria or the weight attributed for the atlas relaxation arise also naturally. At a secondary level, one may also want to investigate the impact of the initialisation of the K-means algorithm in the initialisation of the splitting of a uniform component. With respect to the outlier probabilistic priors, the choice of the outlier probability threshold will lead to differences in the initial classification as outlier used to build the subject-specific outlier atlas (cf Section 3.2.2). Moreover, it must be noted that the classification as inlier or outlier is impacted differently for the various anatomical tissues considered. At a stage where there is no constraint yet on the covariance matrix, the change in inlier/outlier classification will be more important in tissues for which the covariance matrix is larger and therefore display a flatter probability distributions such as the CSF. A variant in the building of the outlier atlas that will be developed in Section 7.2 is to use typicality maps as defined by VanLeemput et al. [79] so that such class-dependent outlierness assessment is avoided. A parameter is also available to tune the conservatism of the outlier definition.

Regarding the values of the MRF energy matrix, lower values will offer more flexibility in the segmentation, that may therefore appear noisier, while higher values will contribute to a smoother segmentation to the expense of a loss in details. Further tuning could involve a different weighting between tissues or even local choices of energy matrix.

For the choice of the MRF energy matrix, a trade off must be found between the spatial regularisation and the lack of detail. In order to behave similarly to the MRF weighting of 0.15 used in the segmentation process of the VBM tool included in SPM8 and also by Ortiz et al. [206], the symmetric matrix H containing the neighbourhood energy cliques for the MRF is defined as:

$$H(l_{jk}, l'_{j'k'}) = \begin{cases} 0 & \text{if } j = j' \\ 0.15 & \text{otherwise} \end{cases}$$

Enforcing spatial consistency only at Level 2 of the hierarchical model should avoid the smoothing of relatively small but very hyperintense lesions that will be detected thanks to the inlier/outlier separation.

Different aspects of BaMoS behaviour are here investigated with respect to different potential modifiers. In Section 5.4.1, the impact of preprocessing choices are evaluated using the ADNI92 dataset. Then, three models of evolution of BaMoS are tested in the simulated cases as well as with the MICCAI MS and the T2DB datasets. Generalisability issues with respect to choice of modalities are then presented in Section 5.4.3 using the simulated Brainweb data and the SABRE dataset. The issue of protocol MR acquisition and image resolution is examined in Section 5.4.4 using the POPPY dataset. Finally the postprocessing stage and its influence on assessment results is discussed in Section 5.4.5. Such evaluations are indeed pivotal to understand the degree of robustness of a given algorithm and assess the uncertainty related to measurements according to the conditions of application.

5.4.1 Impact of preprocessing

In order to investigate the impact of preprocessing choices, three possible modifiers were selected:

Segmentation space Since the model selection process uses different pulse sequences, a common space must be chosen to perform the segmentation and images registered together. Although a manual segmentation is likely to be performed in the native space of the FLAIR image, automatic methods are often applied in the T1 space to which the FLAIR image is registered. Indeed, the T1 image is generally of higher resolution and may be used in a preprocessing step before consideration of the FLAIR image. Note however that the impact of the registration parameters is not investigated here.

Intracranial volume extraction In BaMoS, mask extraction methods may prove an important aspect since it is the basis of the intensity normalisation. Additionally, changes in the brain mask can contribute in different sources of potential false positives and affect the post-processing if this one is based on distance to the border of the mask.

Statistical atlases Statistical atlases that hold *a priori* information on the anatomy are inherent elements of the segmentation process. Naturally, the choice of such information may affect the final outcome.

5.4.1.1 Experiments

For the masking strategy, two options (mentioned in Section 4.3.1) were considered: the first consisted in applying morphological operations (dilation by 2 voxels, filling and erosion by 1 voxel) on the brain delineation obtained as output of BrainMAPS [207] while the second consisted in directly using the total intracranial mask obtained from the pipeline based on the Geodesic Information Flow (GIF) method developed by Cardoso et al. [198]. With respect to the statistical priors, the ICBM priors registered non-rigidly [208] to the studied cases was compared to the tissue priors obtained as fusion of propagated tissue segmentations through GIF. Figure 5.12 illustrates the differences in masks and atlases.

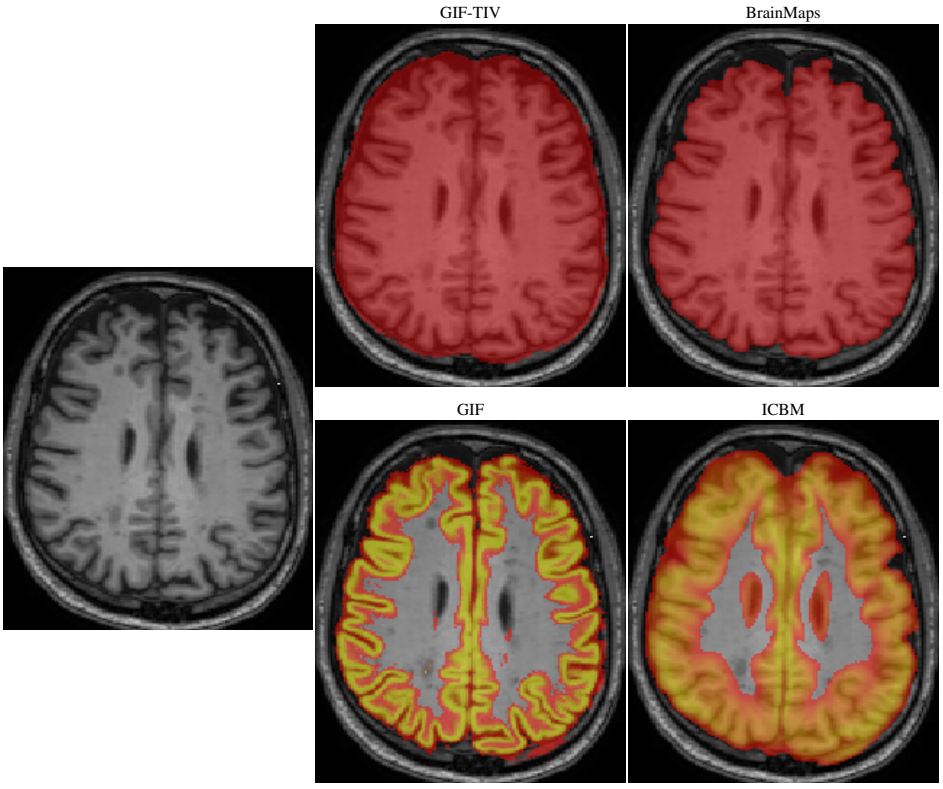


Figure 5.12: The top row presents the masks used overlayed on the T1 weighted image presented on the left with the TIV obtained with GIF on the left and the morphologically modified BrainMAPS on the right. The bottom row presents the GM statistical atlases obtained with GIF on the left and the ICBM template on the right.

	FLAIR to T1				T1 to FLAIR				WMHU
	GIF	**GIF-BM**	**ICBM-BM**	**ICBM**	**GIF**	**GIF-BM**	**ICBM-BM**	**ICBM**	
Mean	5.52	6.37	6.26	5.52	6.91	6.05	6.58	6.93	7.08
SD	6.06	6.59	6.33	6.07	7.39	6.5	6.97	7.28	7.73
Median	3.39	4.16	4.21	3.59	4.12	4.07	4.56	4.35	4.34
IQR	[1.66 6.74]	[1.75 7.87]	[2.03 7.56]	[1.53 6.56]	[1.95 8.79]	[1.58 7.72]	[1.89 8.76]	[2.20 8.26]	[2.42 8.44]

Table 5.4: Volumetric measurements obtained in the different conditions of processing. WMHU indicates the volumes reported online. All volumes are given in mL.

		GIF-BM	**GIF**	**ICBM-BM**	**ICBM**
T1toFLAIR	R^2	0.89	0.92	0.87	0.93
	Slope	0.79	0.92	0.84	0.91
	Cons	0.46	0.46	0.66	0.54
FLAIRtoT1	R^2	0.92	0.91	0.91	0.82
	Slope	0.82	0.751	0.78	0.72
	Cons	0.64	0.24	0.76	0.47

Table 5.5: Regression results obtained when comparing the different measures obtained to the publicly reported WMH volumes.

BaMoS was performed on all combinations of space, masks and atlases resulting in eight measurements for each of the 92 cases. Due to the skewness of the data, Wilcoxon tests were used to compare the medians. Linear regression parameters were also evaluated.

5.4.1.2 Results

Table 5.4 presents the summary of the volumetric measures for the different combination of processing choice. The choice of priors is indicated as ICBM or GIF while the indicator BM is used when the BrainMAPS algorithm is used. Secondly, the regression coefficients between the eight measurements and the measures reported online for the WMH volumes are presented in Table 5.5. In turn, Table 5.6 summarises the comparisons between registration space for the four other parameter combinations. Lastly the pairwise comparisons of the results obtained in both spaces between choice of statistical atlases and choice of intracranial volume definition is presented in Table 5.7.

	GIF-BM	**GIF**	**ICBM-BM**	**ICBM**
R^2	0.93	0.93	0.83	0.82
Slope	0.98	0.79	0.83	0.75
Cons	0.45	0.05	0.80	0.29
p-value	0.12	$<$0.0001	0.7202	$<$0.0001

Table 5.6: Regression results for each set of parameter between the results obtained in the T1 space and the results obtained in the FLAIR space.The p-value refers to the comparison in volumes using a paired Wilcoxon test.

Comparison		TIV		Atlases		Crossed	
		GIF-BM GIF	ICBM ICBM-BM	GIF-BM ICBM-BM	ICBM GIF	GIF-BM ICBM	ICBM-BM GIF
T1toFLAIR	R^2	0.95	0.93	0.88	0.95	0.95	0.91
	Slope	0.86	1.01	0.88	0.96	0.87	0.9
	Cons	145	275	284	286	33	356
	p-value	<0.0001	0.006	0.11	0.55	<0.0001	0.0136
FLAIRtoT1	R^2	0.96	0.87	0.95	0.88	0.86	0.95
	Slope	1.07	0.89	1.01	0.94	1.00	1.02
	Cons	478	-72	32	335	828	633
	p-value	<0.0001	0.0002	0.71	0.12	0.0001	<0.0001

Table 5.7: Comparison between WMH volume measurements when changing intracranial volume extraction method or statistical atlases.

Instead of building a common segmentation, consistency measurements as defined in Section 5.2.4 are used to evaluate the delineations. These results are gathered in Table 5.8. Segmentation results appeared to be heterogenous in the infratentorial region and an example of such variability is presented in Figure 5.13. In order to further highlight the impact of the mask on the normalisation and its consequent effect on the lesion segmentation, Figure 5.14 presents the segmentations and their overlap for each choice of statistical atlas.

5.4.1.3 Discussion

Although it can be meaningful to investigate the absolute differences between measurements performed under different tested and controlled conditions, these comparison do not bring much information on whether or not it is possible to reliably trust one setting of measurements or find reliable ways to compare it to measures performed under other conditions. Such conversions are expressed in the results of linear regressions. Here,

		FLAIR to T1				T1 to FLAIR			
		GIF	GIF BM	ICBM BM	ICBM	GIF	GIF BM	ICBM BM	ICBM
PropWM	Mean	0.12	0.14	0.12	0.14	0.12	0.11	0.13	0.13
	SD	0.03	0.03	0.04	0.04	0.03	0.03	0.04	0.11
	Median	0.12	0.13	0.11	0.12	0.11	0.11	0.13	0.12
	IQR	[0.10 0.14]	[0.11 0.16]	[0.10 0.16]	[0.10 0.17]	[0.10 0.14]	[0.10 0.13]	[0.11 0.16]	[0.10 0.15]
PropLesion	Mean	0.55	0.54	0.57	0.58	0.23	0.52	0.52	0.54
	SD	0.03	0.03	0.04	0.04	0.03	0.03	0.04	0.11
	Median	0.54	0.53	0.56	0.57	0.19	0.50	0.49	0.52
	IQR	[0.41 0.66]	[0.42 0.66]	[0.42 0.70]	[0.46 0.70]	[0.15 0.26]	[0.40 0.63]	[0.39 0.66]	[0.40 0.69]
DistQuant	Mean	2.37	2.37	2.32	2.23	2.54	2.62	2.52	2.39
	SD	0.50	0.50	0.58	0.48	0.50	0.46	0.49	0.87
	Median	2.39	2.40	2.38	2.29	2.50	2.69	2.53	2.53
	IQR	[2.00 2.73]	[2.05 2.69]	[1.97 2.64]	[2.00 2.55]	[2.97 3.75]	[2.97 3.36]	[2.88 4.07]	[2.90 3.65]

Table 5.8: For each of the preprocessing combination, the statistical results of the consistency measurements are summarised.

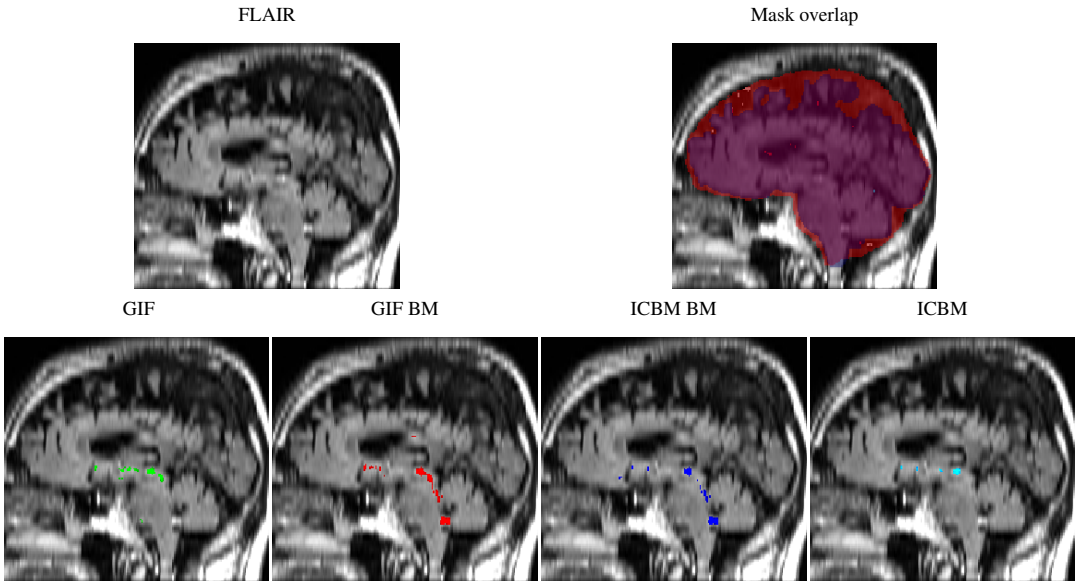


Figure 5.13: On the top row, the FLAIR image on the left and the overlapped masks are presented. In red the TIV mask obtained through GIF and in purple blue the mask obtained after morphological operations on the BrainMaps results. On the bottom row are presented the four segmentations according to choices of atlases and masks.

the following conclusions can be drawn from the combinations of the three choices.

With respect to the space of reference, a high correlation (R^2=0.93) was observed when using GIF atlases whereas the use of ICBM atlases led to a slightly reduced value (R^2=0.83). Furthermore, the volume detected in the FLAIR space was higher than when segmented in the T1 space and this difference was significant when using the TIV directly obtained with GIF. This difference could be partly related to the lower resolution of the FLAIR images for which each individual voxel corresponds to a larger volume. Additionally, the results for the difference between the third quartile of the healthy white matter and the first quartile of the lesion distribution appeared to be higher when segmenting in the FLAIR space than in the T1 space in all other considered conditions. Reasons for such a systematic difference may implicate the influence of partial voluming introduced when resampling the 2D FLAIR image onto the higher resolution isotropic T1 image thereby producing fuzzier lesion boundaries. As exposed on Figure 5.14, most of the difference across masks and prior choices occurred at the border of lesions. It must also be noted that the space of reference especially when associated to a change in resolution may also affect the masks and statistical priors themselves. For both ICBM and GIF priors, the application of morphological changes

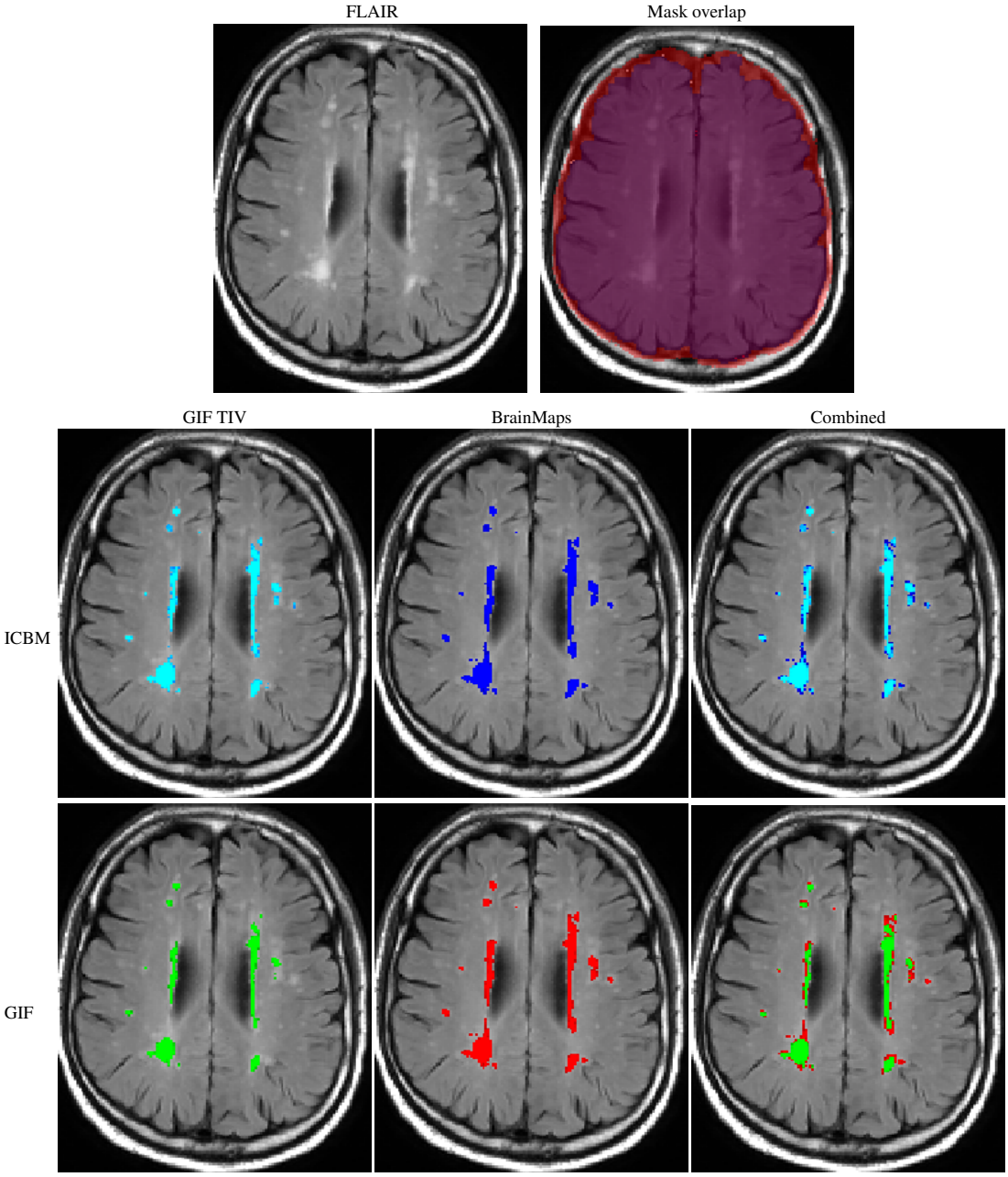


Figure 5.14: On the top row, the FLAIR image and the overlapped masks are presented. Again, the TIV mask in red encompass all the blue brainmask morphologically transform from the BrainMaps output. The middle row presents the segmentations obtained for each mask when applying the ICBM priors and their combination while the bottom presents the segmentations when the priors obtained through GIF are used.

on the BrainMAPS output led to a stronger additive bias than when using the TIV output of GIF.

When studying the pairwise comparisons within each reference space, one can observe that when using GIF atlases, the strategy of TIV extraction affected significantly the volume obtained but the correlation was very high (R^2=0.95 and 0.96 according to the segmentation space), thus reflecting directly the variation in intensity normalisation mentioned earlier. The choice of the atlases appeared to have a different impact according to TIV extraction method and segmentation space chosen. The impact of the choice of priors was made visible through the variability of the consistency measures; in all cases the standard deviation of the measures were higher when using the ICBM atlases compared to the GIF tissue priors. Such difference could be interpreted as the consequence of the sharpness of the used atlases in the case of GIF, potentially reducing the variation due to registration issues that may occur with varying anatomies.

Using as a reference the volumes of WMH reported online, the GIF statistical atlases appeared to yield higher correlations. Caution must however be applied when using such reference volumes since they can themselves be subject to errors and performance variability.

5.4.2 Impact of model evolution

5.4.2.1 Experiments

In order to assess the impact of the model selection process and the importance of the model's adaptability, BaMoS was compared to two simpler versions of itself in which the automatic selection of lesion is implemented. The following variants of BaMoS were used for comparison: the first one, called BaMoS-static consists in the static solution obtained without allowing for a change in the number of model components, *i.e.* performing the first initial EM, the atlas adaptation and a final EM refinement. The second simpler version of BaMoS, denoted BaMoS-NoCov is the one detailed in the work presented at the MICCAI 2014 conference [196], *i.e.* without any constraint on the covariances. Finally the final version of BaMoS is simply denoted BaMoS. Both the simulated Brainweb data and the T2DB dataset were used to assess the performance of the three versions.

		Assessment method							
		DSC	VD	FPR	TPR	FNR	AvDist	DE	OER
Mild	BaMoS-NoCov	36.9	94.3	62.9	44.2	55.8	8.3	84.3	101.7
	BaMoS	**51.4**	46.4	44.9	52.0	48.0	**2.7**	**40.9**	86.5
	BaMoS-static	48.6	**43.1**	**25.6**	42.9	57.1	4.5	45.8	74.5
Moderate	BaMoS-NoCov	55.7	75.3	91.6	72.1	27.9	1.3	507.6	104.8
	BaMoS	**70.2**	**22.9**	33.6	**72.8**	27.2	**0.8**	**173.9**	55.3
	BaMoS-static	65.0	33.9	**20.8**	60.2	39.8	1.3	309.2	51.3
Severe	BaMoS-NoCov	77.2	37.5	46.3	**91.2**	**8.8**	0.5	71.6	54.3
	BaMoS	**81.2**	23.2	32.6	90.5	9.5	**0.4**	**33.6**	41.7
	BaMoS-static	79.7	**20.1**	**16.4**	77.9	22.1	**0.4**	50.1	**38.0**

Table 5.9: Comparison for the different assessment measures of the segmentation method for the T1T2 combination modality. The results are taken as the mean over all level noise and IIH at the three different lesion loads.

5.4.2.2 Results on simulated data

The various lesion segmentation assessment measures for T1/T2 input data are gathered in Table 5.9 across noise and IIH level and for each lesion load.

The robustness to the noise level for the different lesion loads is presented using the DSC in Figure 5.15 for which the methods are applied on the T1T2 modality combination.

At each noise level, the errorbars present the minimum and maximum result obtained when varying the intensity inhomogeneity level from 0 to 40% by steps of 20%.

Table 5.10 congregates the assessments at various noise levels, supporting the notion that BaMoS is robust to this effect.

When increasing the noise level, a decrease in FP was observed. As shown by the

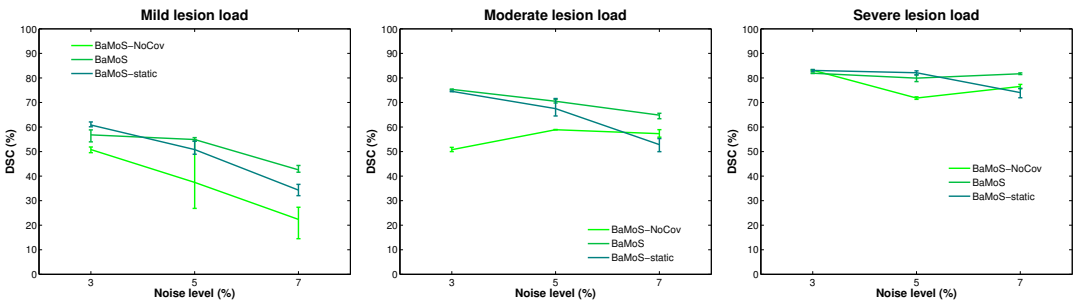


Figure 5.15: Comparison of DSC results for the automated methods with noise level variation at mild (left), moderate (middle) and severe (right) lesion load. The errorbars refer to the minimum and maximum obtained when varying the intensity inhomogeneity level.

Load	Noise	DSC	VD	FPR	TPR	FNR	AvDist	DE	OER
Mild	3	56.8	64.9	54.3	75.4	24.6	1.3	42.7	108.4
	5	54.9	17.7	39.1	50.0	50.0	2.1	32.0	77.5
	7	42.6	56.6	29.7	30.5	69.5	4.7	48.0	73.7
	Mean	51.4	46.4	41.1	52.0	48.0	2.7	40.9	86.5
Moderate	3	75.3	10.3	28.2	79.2	20.8	0.5	44.0	50.1
	5	70.5	33.5	38.3	82.3	17.7	0.6	59.3	66.9
	7	64.8	24.9	24.3	56.8	43.2	1.4	418.3	49.1
	Mean	70.2	22.9	30.3	72.8	27.2	0.8	173.9	55.3
Severe	3	81.9	28.7	27.2	93.7	6.3	0.4	45.3	40.9
	5	79.9	34.6	30.3	93.7	6.3	0.4	30.0	46.8
	7	81.7	6.2	20.7	84.2	15.8	0.4	25.3	37.4
	Mean	81.2	23.2	26.1	90.5	9.5	0.4	33.6	41.7

Table 5.10: Assessment of BaMoS for various measures at different noise level and lesion load for the T1T2 modality combination (mean over various intensity inhomogeneity levels). AvDist is given in mm, DE in μL and all other measures in %.

related changes in DE and OER, misdetection of lesions increased only for the change between 5 and 7% noise on the moderate case. The number of subclasses necessary to model the data was negatively correlated with the noise level, as the wider class variance makes it hard to justify the need to use more Gaussian classes under the BIC model.

5.4.2.3 Results on clinical data

On the MICCAI MS dataset, when comparing the three versions of BaMoS, the only significant differences observed were related to the DSC and the TPR for which BaMoS performed significantly better than BaMoS-static.

The comparison between the three versions of BaMoS on the T2DB dataset, is presented in Figure 5.16. The apparently surprising observation that BaMoS-static did not contain any DSC of 0 is caused mostly by the fact that the lesion segmentation is performed on a voxelwise basis for this method. The only DSC of value 0 observed for BaMoS corresponds to the case where only 24 voxels were manually segmented as lesion. Overall, BaMoS performed significantly better than BaMoS-static for global assessment measures (DSC, VD, TPR) but no significant difference was observed for the Average distance, DE and OER. BaMoS-NoCov performed significantly better than BaMoS-static for VD, TPR but not for DSC. BaMoS-static performed significantly

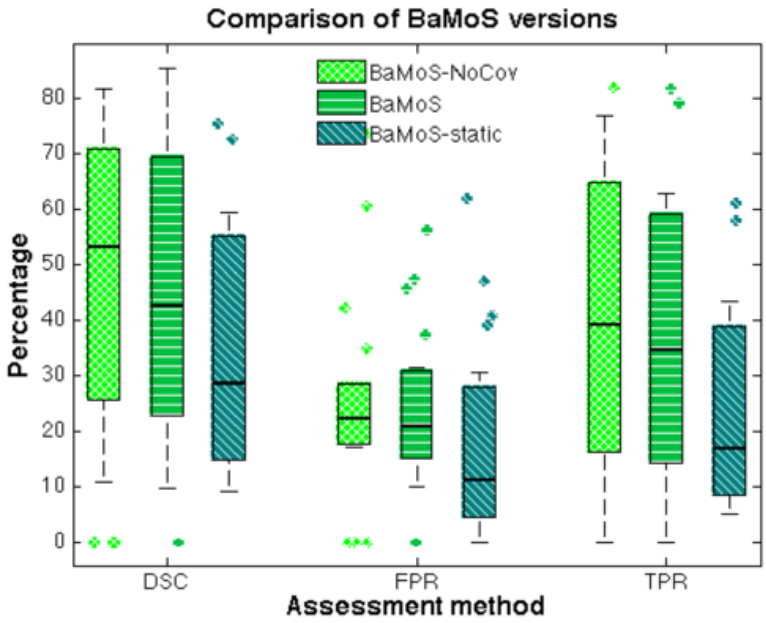


Figure 5.16: Comparison of the three versions of BaMoS in terms of DSC, FPR and TPR. Note the existence of low outliers for the DSC in the BaMoS-NoCov version. The only outlier for BaMoS corresponds to the case with only 0.06 mL of lesion.

better in terms of FPR compared to BaMoS and BaMoS-NoCov but this observation is directly linked to the low TPR. The seeming discrepancy in results showing a higher median for DSC and TPR of BaMoS-NoCov compared to BaMoS, but a significant improvement of the DSC results of only BaMoS over BaMoS-static is related to the presence of outliers in the results of BaMoS-NoCov. This outlines the positive impact on the robustness of the method when including the prior over the covariance.

5.4.2.4 Discussion

Comparing three versions of the proposed methodology: BaMoS-static, BaMoS-NoCov and BaMoS in its full version, BaMoS was observed to be more robust to IIH than BaMoS-NoCov and more robust to noise than both BaMoS-static and BaMoS-NoCov. BaMoS-static appeared to perform reasonably well as the Brainweb data can be modelled by a limited number of Gaussian components. However, the TPR was clearly lower for BaMoS-static than for the other BaMoS versions. The constraint over the covariance matrix for the lesions contributed to promote smaller covariances and thus encourage a more detailed characterisation of difference in levels of lesion severity and partial volume effect.

When comparing the subversions of BaMoS for the clinical data with the eight

different assessment measures presented in Table 5.3, the static version of BaMoS appeared to be less sensitive to the presence of lesion. BaMoS performed significantly better than BaMoS-static in terms of DSC for both clinical datasets. This tends to illustrate the improvements brought by the use of a higher number of Gaussian components in the data modelling. In turn, despite performing marginally worse than BaMoS-NoCov in terms of median (BaMoS 42.6% *vs* BaMoS-NoCov 53.2%), BaMoS was found to be more stable as shown in Figure 5.16 (illustrated by a higher mean BaMoS 46.2% *vs* BaMoS-NoCov 45.2%). The lower sensitivity to lesion in BaMoS-static expressed itself both in terms of a significantly lower TPR and a significantly better FPR when compared to BaMoS and BaMoS-NoCov for the T2DB dataset. The higher robustness of BaMoS compared to BaMoS-NoCov was further exemplified in the case of the clinical MS dataset for which significant differences between subversions was only observed for BaMoS which performed significantly better than BaMoS-static in terms of DSC, FNR and TPR.

5.4.3 Impact of available modalities

5.4.3.1 Simulated data

Applied to the Brainweb data, the graphs in Figure 5.17 show the impact of the noise level on the result for BaMoS at various lesion loads for the different modality combinations, while Table 5.11 gathers the mean DSC across noise and IIH level for the compared versions performed on those combinations.

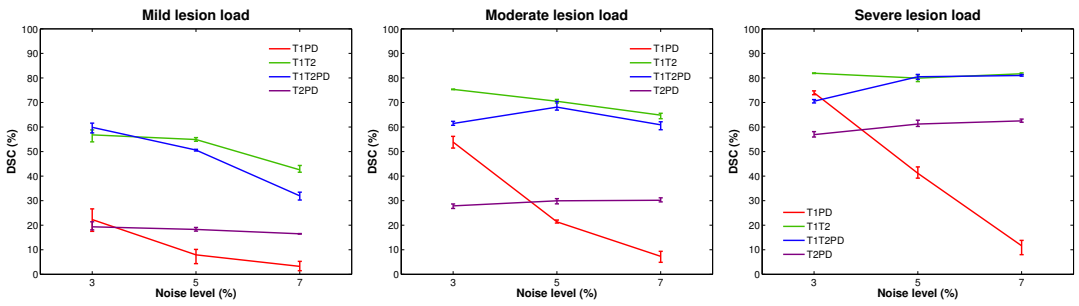


Figure 5.17: Comparison of the DSC results for BaMoS at different noise levels for different modalities combinations for mild (a), moderate (b) and severe (c) lesion load. The errorbars indicate the minimum and maximum obtained when varying the IIH level.

		Modality combination			
Load	Method	T1PD	T1T2	T1T2PD	T2PD
Mild	BaMoS-NoCov	10.3	36.9	35.8	17.9
	BaMoS	11.1	51.4	47.5	18.0
	BaMoS-static	10.0	48.6	40.9	16.7
Moderate	BaMoS-NoCov	29.1	55.7	56.7	28.4
	BaMoS	27.5	70.2	63.5	29.3
	BaMoS-static	23.8	65.0	59.9	32.2
Severe	BaMoS-NoCov	40.0	77.2	72.3	60.6
	BaMoS	42.2	81.2	77.3	60.2
	BaMoS-static	32.8	79.7	77.8	57.9

Table 5.11: Mean DSC (%) over noise and IIH level for the different modality combinations across the different lesion loads for the three compared versions of BaMoS.

5.4.3.2 Clinical data

In this experiment about the behaviour of BaMoS in various conditions of modality choice, the effect of the inclusion of the information coming from a T2-weighted image was investigated applying this time BaMoS to the SABRE dataset described in Section 5.3. In BaMoS, when presenting from the start three modalities, the high degree of freedom in the evolution may lead to biologically implausible models. In order to constrain the model evolution, a third modality is incorporated sequentially to the model: after an initial optimisation for two modalities, the obtained model is used as initialisation point when including the third modality. Alternatively, one may have further enforce the shape of covariance matrices or derive tissue-specific complexity criteria but these avenues were not investigated.

In a first stage, the segmentations obtained with 2 (T1FLAIR) and 3 modalities (T1FLAIR + T2) are compared. Secondly, the order in which the modalities are incorporated (T1FLAIR + T2) or (T1T2 + FLAIR) is evaluated. As in the previous section, volumetric linear regressions are performed on the log-transformed data. The segmentation comparisons were performed using the T1FLAIR+T2 segmentation as reference. The total volumes detected in the three tested cases are presented in Table 5.12.

With respect to the inclusion or not of the T2 modality, the raw linear regression over total lesion volumes led to a R^2 value of 0.97 while the R^2 on log-transformed volumes was of 0.88. The plot of the linear regression on the log-transformed volumes is presented in Figure 5.18 left while the relationship between DSC and volumes of

	T1T2+FLAIR	T1FLAIR+T2	T1FLAIR
Mean	3.42	3.87	3.83
SD	4.30	4.72	4.72
Range	[0.32 21.84]	[0.38 25.28]	[0.13 24.28]
Median	1.38	1.71	2.17
IQR	[0.75 4.46]	[0.83 4.61]	[0.75 4.72]

Table 5.12: Statistics of volumes for the different modality choices. All volumes are presented in mL.

WMH is displayed in Figure 5.18 right.

An illustration of the discrepancies between the two segmentations in the frontal areas is presented in Figure 5.19.

When comparing the order of modality inclusion, the R^2 of the regression between global volumes was of 0.97 and the one on the log-transformed volumes of 0.87. The linear regression between log-transformed volumes is plotted in Figure 5.20 left and the relationship between DSC and volumes in Figure 5.20 right.

Periventricular frontal areas were more prone to disagreements between segmentations. These regions do have high priors to belong to GM due to their proximity to basal ganglia regions and voxels with a limited outlierness may or not be classified as lesions. But it is mostly in the infratentorial regions, regions for which the FLAIR acquisition is known to provide less reliable signal, that more WMH are detected when using the FLAIR image first. Such a behaviour is exposed in Figure 5.21.

As for the measures of similarities, Table 5.13 summarises the statistics for the

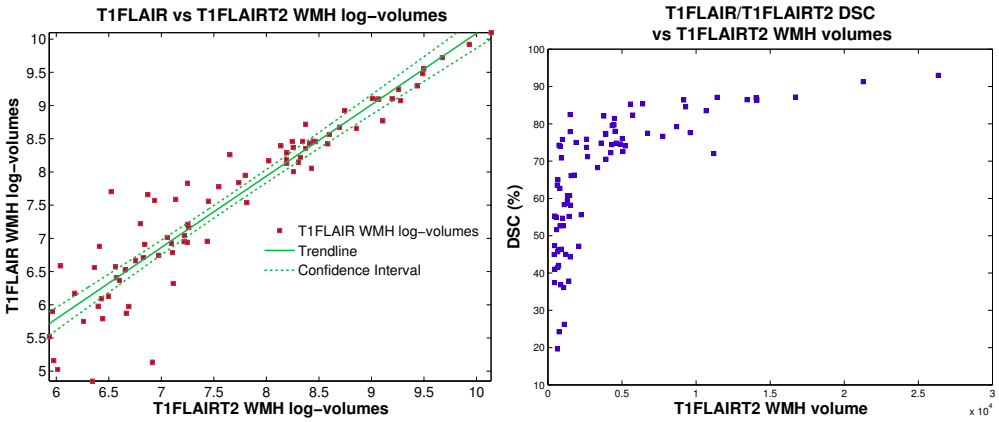


Figure 5.18: Left) Global log-transformed volumetric linear regression between T1FLAIR and T1FLAIRT2 segmentations Right) Relationship between DSC using the T1FLAIR+T2 as reference and the WMH volumes of this reference segmentation.

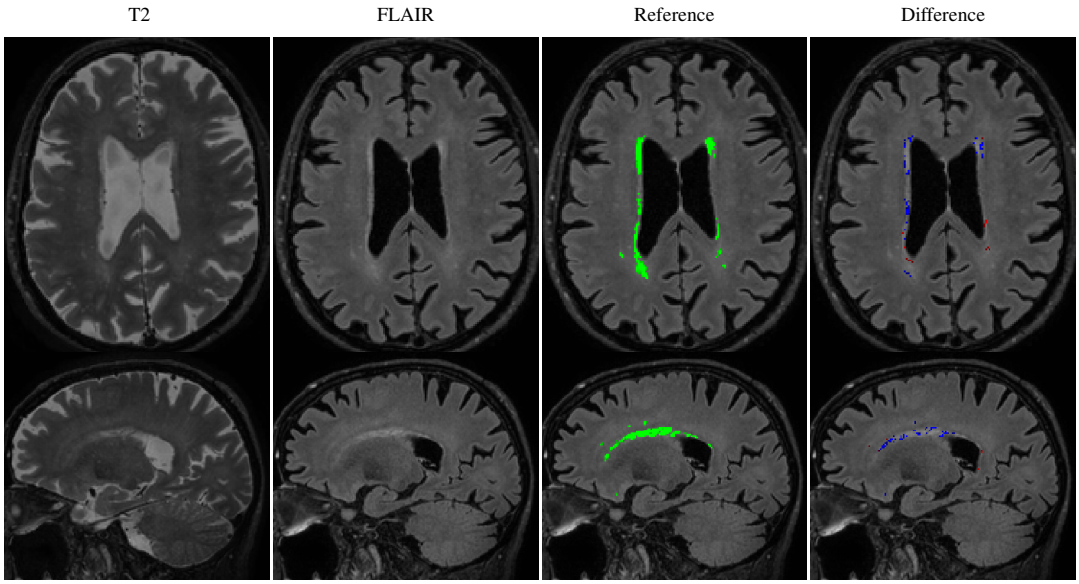


Figure 5.19: Example of discrepancies of segmentation in the frontal areas when using 2 or 3 modalities. The reference segmentation is the one using T1FLAIR+T2 modalities while the difference is obtained by (T1FLAIR)-(T1FLAIR+T2). The blue voxels correspond to the false negatives with respect to the reference image *i.e* the voxels considered as lesion when using the three modalities but not when using only T1 and FLAIR while red voxels are the false positives.

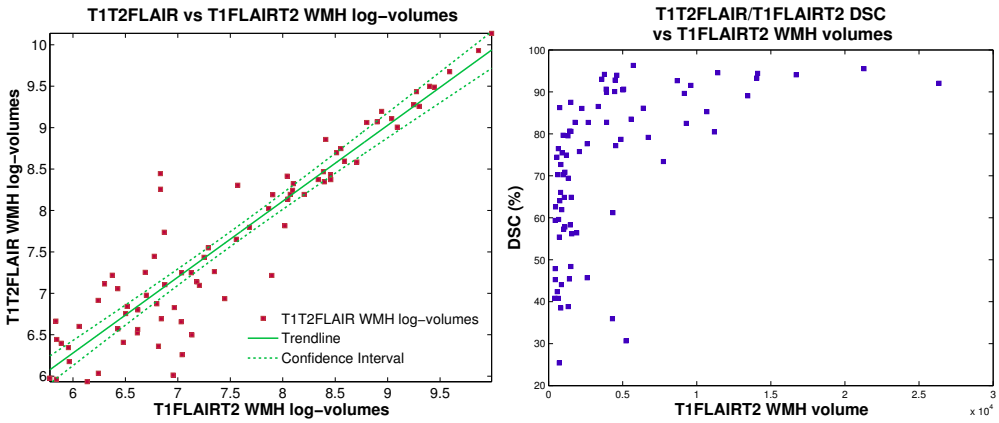


Figure 5.20: Left) Global log-transformed volumetric linear regression between T1FLAIR+T2 and T1T2+FLAIR segmentations. Right) Relationship between DSC of the T1T2+FLAIR segmentation compared to the T1FLAIR+T2 segmentation and the volume of the T1FLAIR+T2 segmentation.

		DSC	AvDist	TPR	FPR	OEFP/FP	OEFN/FN	OE/TotF	FP/TotF
	Mean	65.5	1.98	66.1	47.3	70.7	65.2	67.3	47.3
	SD	17.3	1.91	19.4	20.8	13.0	17.8	14.5	20.8
T1FLAIR	**Range**	[19.6 93]	[0.19 8.99]	[12.2 92.1]	[2.3 87.5]	[19.4 90.6]	[23 90.8]	[34.3 88.9]	[2.3 87.5]
	Median	71.6	1.20	71.4	48.8	71.9	70.6	70.3	48.8
	IQR	[52.7 77.9]	[0.63 2.93]	[54.3 81]	[31.8 63.9]	[64.1 81.1]	[47.5 79]	[56.1 79.8]	[31.8 63.9]
	Mean	72.2	1.86	69.8	20.2	66.4	57.8	62.5	35.7
	SD	18.6	1.88	21.4	18.3	15.9	19.2	15.9	25.9
T1T2+FLAIR	**Range**	[25.4 96.3]	[0.1 8.76]	[18.3 97.5]	[0.6 68.9]	[36.2 93.6]	[12.3 87.1]	[29.1 86.4]	[1.4 93.4]
	Median	76.8	1.21	75.4	14.1	68.2	60.8	64.0	29.1
	IQR	[58.6 88.7]	[0.45 2.62]	[58.4 87.8]	[4.5 30.6]	[52.7 79.3]	[40.3 73.6]	[47.9 76.1]	[13.2 57.0]

Table 5.13: Summary of the similarity metrics and origin of errors when comparing the choice and the order of the chosen modalities. AvDist is given in mm while all other measures are given in %.

different comparison assessments taking as reference the segmentation obtained using T1FLAIR+T2 modalities.

With a mean DSC above 65% in both situations, the agreement between segmentations can be judged as excellent. In both cases, most of the differences reflected a disagreement in the definition of the lesion borders with an outline error percentage of 67.4 when using only T1FLAIR and 62.5 when employing the T2 modality first.

5.4.3.3 Discussion

The very strong correlation (R^2=0.97 on raw volumes, 0.88 on log-transformed volumes) between volumetric assessment of lesion when using T1, FLAIR and T2 or not shows the stability of the algorithm when complementing the data information with an additional modality. This strong agreement is reassuring in two ways. First it shows the consistency in the extracted information and second it provides a reliable way of transferring a measurement performed with one set of modalities to another. The disagreement was concentrated in regions where tissues with intensities close to the normal GM are likely to occur and may or not provide the satisfactory criteria to contribute

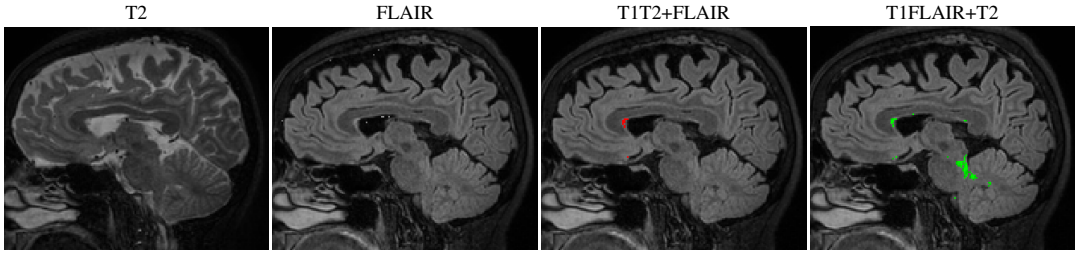


Figure 5.21: Illustration of differences in segmentation according to the order in which modalities are considered for the infratentorial regions. The slight hyperintensities in the FLAIR image are classified as lesion when using the FLAIR (T1FLAIR+T2) before the T2 modality (T1T2+FLAIR).

to the initial outlier maps such as the periventricular frontal regions. A slight shift in the model parameters due to the inclusion of the T2 modality may then transform slightly the GM inlier/outlier separation in those regions for which GM priors are usually high. The use of three modalities contributes to a higher volume of WMH detection in those areas. In contrast, more WMH is detected in the mid zone of the frontal region when using only two modalities. Knowing that most of the FP are coming from the outline of the lesions (71%), this can be explained by the fact that FLAIR images tend to overestimate the extent of lesions compared to T2 images. The excellent overall agreement between results using two (T1FLAIR) or three (T1FLAIR+T2) modalities guarantees that if the T2 modality is missing from one dataset, this one can still be processed and the resultant WMH segmentation used with confidence. Regarding the order in which the modalities are introduced, performing first the model optimisation on the T1T2 couple before incorporating the FLAIR or including the FLAIR modality first also led to a strong agreement between results both on raw (R^2=0.97) and on log-transformed volumes (R^2=0.87). Frontal periventricular and infratentorial regions appeared as before to be more prone to disagreements than other brain areas.

5.4.4 Impact of acquisition resolution

5.4.4.1 Experiment

Using the POPPY dataset, the lesion segmentation was performed for each subject using T1 and the 2D FLAIR (FLAIR2) images and the T1-weighted and the 3D acquired FLAIR (FLAIR3) images. In order to understand relationship and differences, linear regressions between extracted volumes of lesions were performed. Additionally, the two segmentations were compared to examine the source of differences using the FLAIR3 segmentation as reference.

5.4.4.2 Results

The lesion volumes segmented using the 2D acquisition led to a median volume of 2.09 mL (IQR =[1.28 3.49]) ranging from 0.19 to 3.24 mL while the 3D acquisition had a median of 0.84 mL (IQR=[0.37 1.91]) with a range from 0.13 to 2.17 mL . Given the high skewness of the data, regressions and correlations were estimated on the log-transformed data. The overall R^2 between the two obtained volumes was of 0.70 for the log-transformed data and 0.94 for the raw data. Figure 5.22 left displays the regression

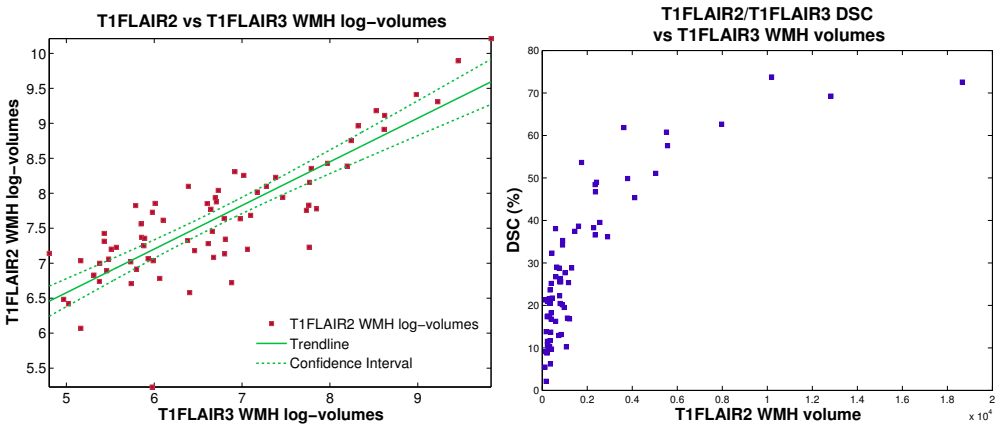


Figure 5.22: Left) Linear regression between log-transformed WMH volumes segmented on 2D and 3D acquired FLAIR images. Right) Relationship between DSC and volume of FLAIR 3 segmentation.

between log-volumes and Figure 5.22 right the relationship between DSC and volumes.

As far as similarity measures are concerned, the DSC followed the pattern with large variability for low volumes with commonly lower values and a plateauing effect at larger volumes as illustrated in Figure 5.22 right. Due to the very low volume of WMH in this dataset and the known relationship between the accuracy measure and the lesion volume, the DSC and TPR appear to be lower than normal as presented in Table 5.14.

Upon visual inspection of the areas of greatest difference, multiple explanations can be put forward: compared to the 3D acquisition, FLAIR2 images presented more artefacts around the fourth ventricle and the Sylvian aqueduct which despite the false positive correction step resulted in the observed segmentation difference in the posterior fossa. This observation is in accordance with reports of the reduction of artefacts in this area in 3D acquired FLAIR compared to 2D images [160, 161, 209]. Figure 5.23 presents the comparison of data and lesion segmentation for such artefacts.

	DSC	AvDist	TPR	FPR	OEFP/FP	OEFN/FN	OE/TotF
Mean	27.8	6.07	46.3	77.8	54.1	60.3	55.4
SD	17.5	3.80	18.6	16.5	19.8	14.2	17.6
Range	[2.1 73.8]	[0.69 18.09]	[8.0 89.2]	[2.9 98.8]	[7.2 97.3]	[30.8 84]	[10.5 94.9]
Median	22.0	5.54	44.9	83.0	51.9	61.0	54.0
IQR	[13.7 37.9]	[2.98 7.83]	[31.2 60.0]	[70.5 90.4]	[40.9 65.2]	[49.6 71.9]	[43.6 68.0]

Table 5.14: Comparison assessments between segmentations taking the FLAIR3 as reference. AvDist is given in mm and all other measures in %.

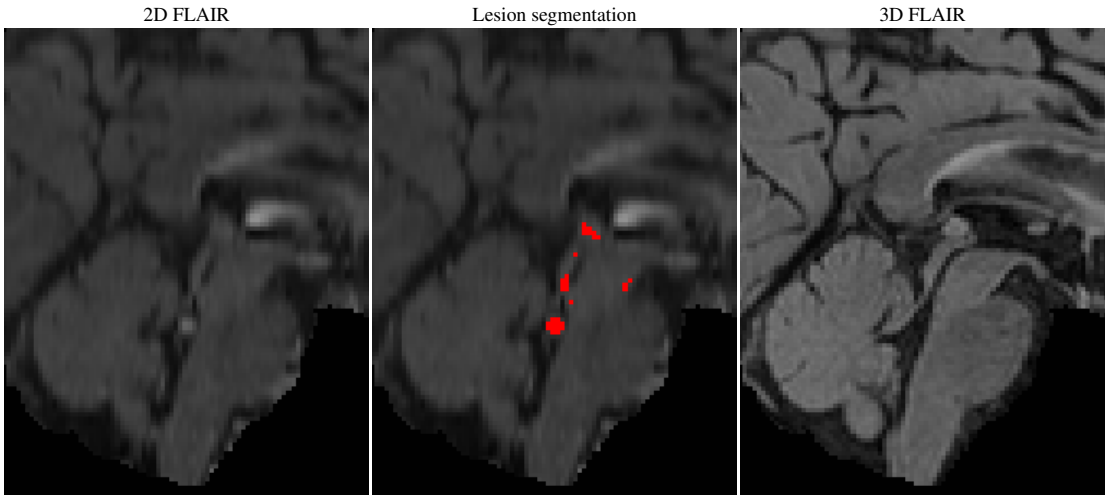


Figure 5.23: Occurrence of flow artefacts with the 2D acquisition (left) segmented as lesion (middle) and the corresponding 3D acquisition.

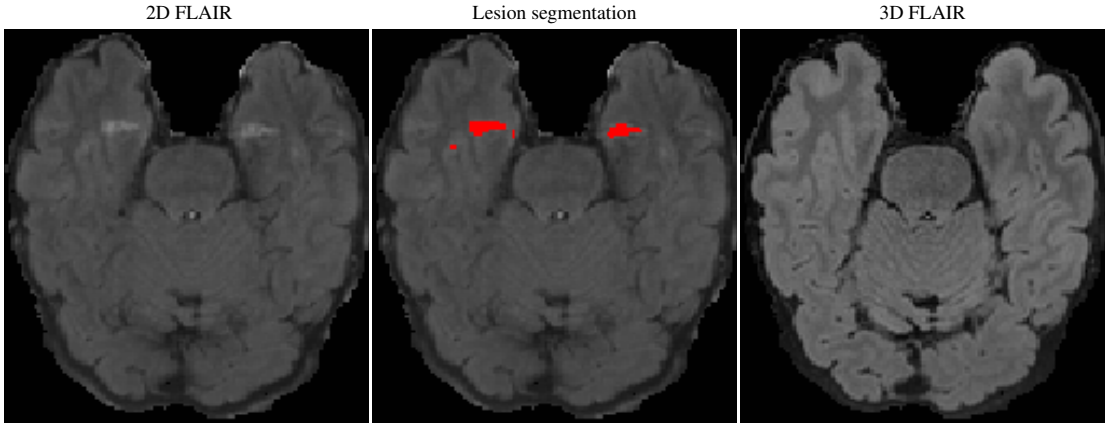


Figure 5.24: Occurrence of artefacts in the temporal lobe with the 2D acquisition (left) segmented as lesion (middle) and the corresponding 3D acquisition.

Other types of artefacts may arise as shown in the temporal lobes in Figure 5.24.

However, 3D acquisition can also be prone to artefacts such as a strong bias field that even corrected may prevent the detection of some lesions as presented in Figure 5.25.

Additionally, close to vessels, flow artefacts may appear in 2D acquisition producing mimics of small lesions, whereas other small lesions about the size of the slice thickness are blurred or invisible in 2D but well defined in the 3D acquisition. Examples of these two situations are presented in Figure 5.26.

When comparing the two acquisition protocols, the lesions acquired in a 2D manner appeared fuzzier with smoother and more extended boundaries. Due to the variation in pulse sequence and acquisition parameters, contrast between tissues varied from one

acquisition to the other. In this example, the contrast between GM and WM appeared stronger in the 3D than the 2D acquisition while the contrast between lesions and GM seemed stronger in the 2D acquisition. As a consequence, the normalisation of intensities is also altered and the behaviour of the algorithm is affected in its decision on the level of outlierness of the tissues. The impact on lesion segmentation is presented in Figure 5.27.

Such differences in classification are most important in regions in which the *a priori* probability to belong to GM is high as may occur close to the ventricles or in the occipital region. To be classified as outlier, the voxels have then mostly to be outliers to the GM distribution and not only to the WM.

5.4.4.3 Discussion

In this experiment, a good agreement was observed between log-transformed total lesion volume obtained by segmentation of 2D or 3D acquired FLAIR images registered in the space of an isotropic T1-weighted image with a R^2 regression coefficient of 0.7. Compared to the evaluations performed in Schmidt et al [90] in which the 3D FLAIR images were downsampled to simulate images with larger slice thickness, differences in scanning protocol led to noticeable changes in image appearance with tissue contrast differences and occurrence of different types of artefacts. If most of the flow and

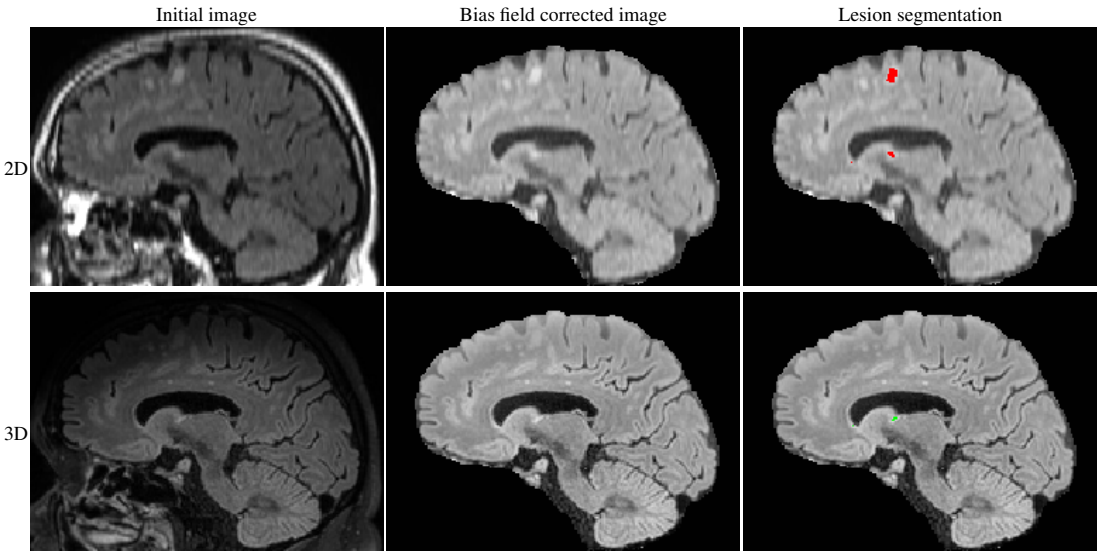


Figure 5.25: Initial image (left), skull-stripped, log-transformed, normalised and bias field corrected (middle) and resulting segmentation (right) for the 2D (top row) and 3D (bottom row) acquisitions.

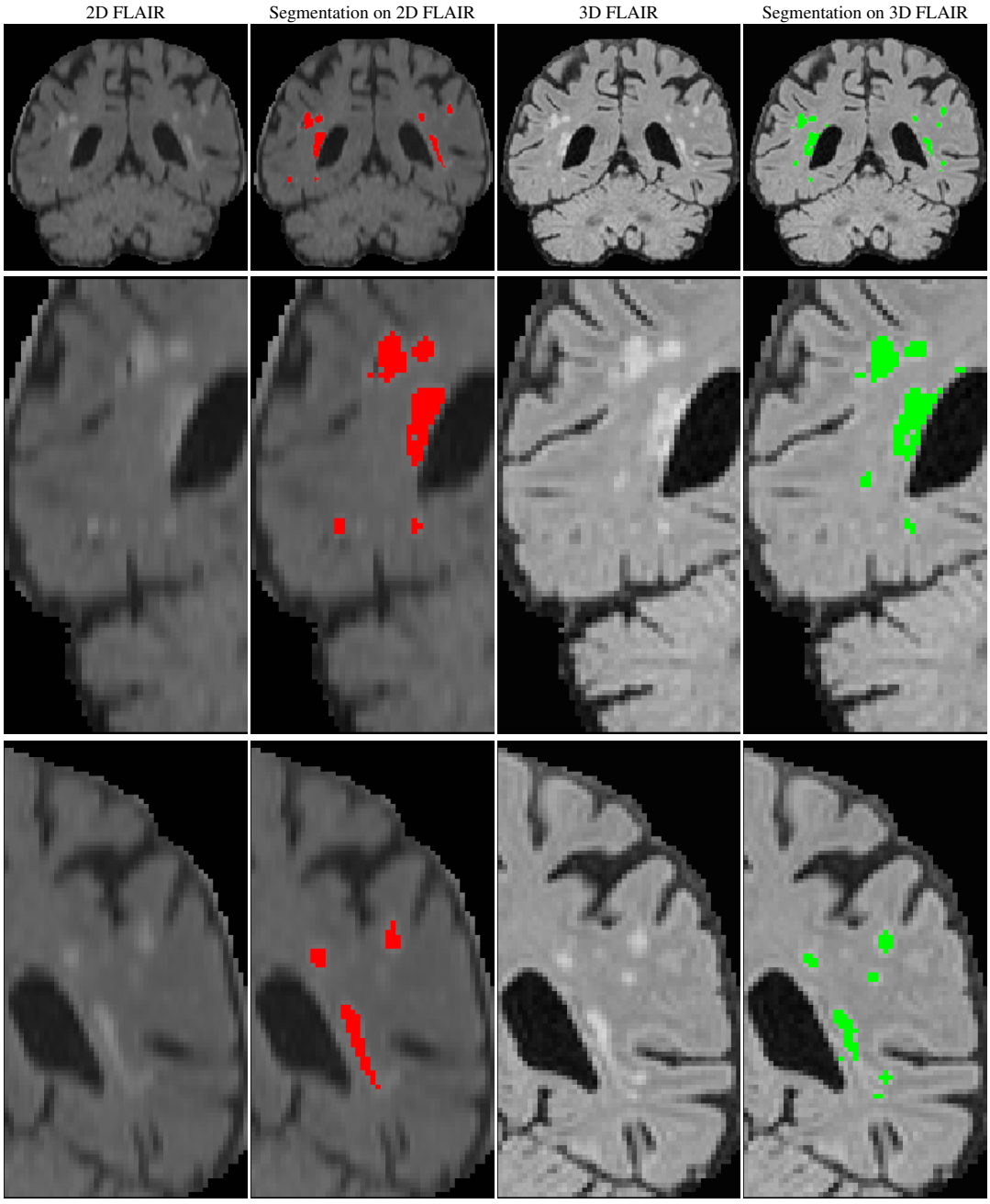


Figure 5.26: On the three rows, the 2D (resp 3D) acquisition and resulting segmentation are presented on the left (resp right) side of the figure. The top row presents the complete slice while the middle row presents a case of artefact in the 2D acquisition absent of the 3D and the bottom row zooms on a small lesion made invisible in the 2D acquisition.

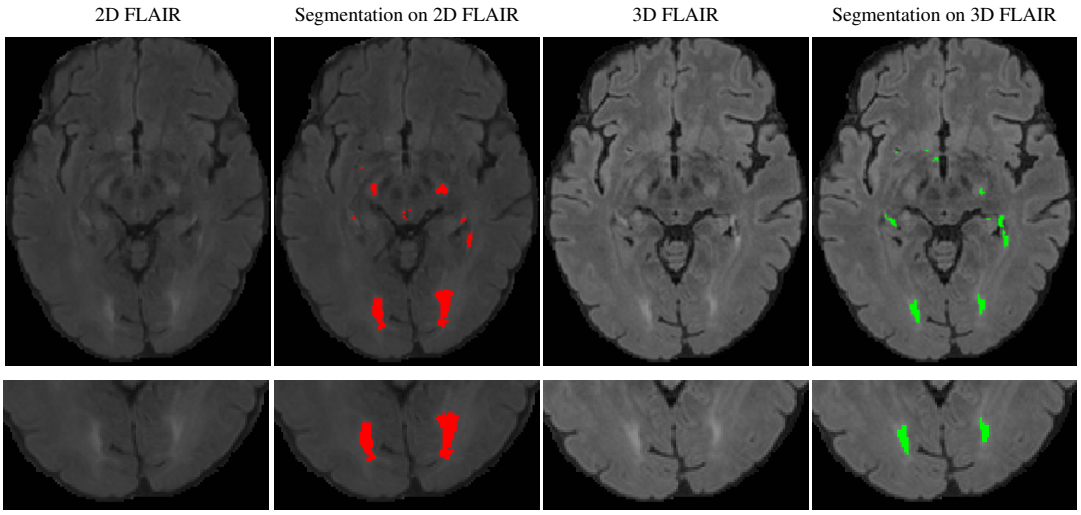


Figure 5.27: Differences in segmentation due to change in contrast and blurriness with the 2D (resp 3D) acquisition and the corresponding lesion segmentation on the left (resp. right).

pulsation artefacts were present with greater strength in the 2D acquisition as reported by Naganawa et al. [161] and Lummel et al. [160], the effect of the bias field appeared stronger in some examples for the 3D acquisition and led to the misclassification of lesions as normal tissues. Although highly correlated, the differences in contrast may justify a different choice in tuning parameters across acquisitions. Practically, this relates to the use of multi-centre data, accounting for scanner upgrade or changes in protocol. Per site/acquisition covariates can be used to increase the number of measurements, as volumes were found to correlate up to a linear correction factor. One may also consider discarding local parts of the information in order to ensure a more reliable correspondence. Such analyses would enable the use of complementary or longitudinal datasets in which clinical relationships could be better highlighted despite a slight increase in measurement uncertainty.

5.4.5 Impact of postprocessing

In order to assess the importance of the postprocessing stage, BaMoS and its non corrected version (BaMoS-nc) were performed on the T2DB and the obtained segmentations were assessed against the available manual segmentation. Figure 5.28 presents the results comparison for the DSC, the TPR and the FPR. The stability in TPR supports the appropriateness of the FP detection.

Despite the carefulness with which the FP detection detailed in Section 5.1.3 has

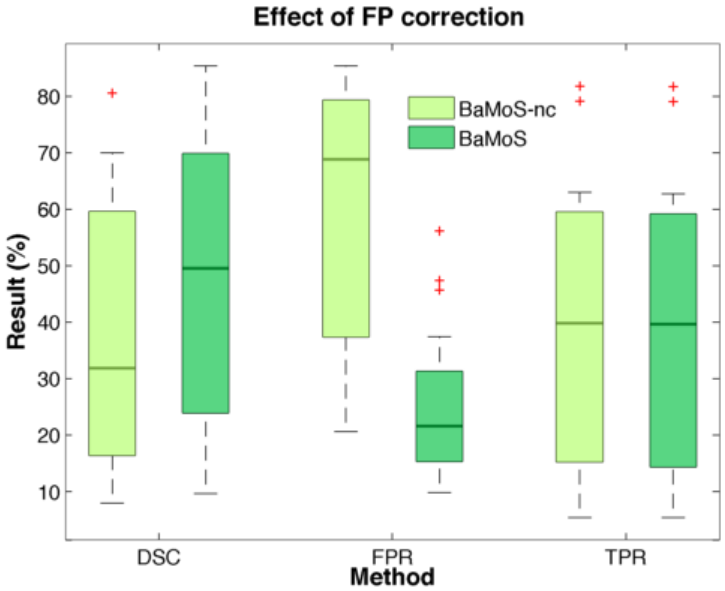


Figure 5.28: Effect of the FP correction on BaMoS in terms of DSC, FPR and TPR. The correction reduces the FPR but does not affect the TPR. BaMoS-nc refers to the result of BaMoS uncorrected for FP.

been made, remaining minor errors were observed in the cortical sheet. Additionally, some FP were also found at the border of some lesions especially for hyperintensies on the FLAIR modality presenting a certain hypointensity on the T1 image that were not considered as lesion in the manual segmentation.

In many methods specially dedicated to the segmentation of white matter lesions, a post processing step is needed to avoid taking into account the hyperintense voxels due to flow artefacts in FLAIR images [94, 210] or voxels at the border between GM and WM [119], or other types of false positives [125]. The FP correction in this work is adapted according to the lesion selection rules. However, it must be underlined that the false positive correction performed here is based on the connected components and is not processed at a voxelwise level which may lead to discard or accept areas that are a mixing of lesion and artefactual voxels.

5.5 External comparison

On top of all internal consistency assessments, the performance of BaMoS was also compared to other families of algorithms in simulated cases using the Brainweb simulated database, and two clinical datasets: MS pathology (MICCAI MS) and age-related WMH (T2DB).

5.5.1 Points of comparison

The competing automated algorithms for lesion segmentation were selected if the corresponding software was available online and had minimal preprocessing requirements (*e.g* skull-stripping). The first one was the classic EMS algorithm [79] that belongs to the same family of methods as BaMoS and thus enables a very similar set-up. The EMS code, available online (`https://mirc.uzleuven.be/MedicalImageComputing/downloads/ems.php?EMSsection=download&pagePath=2`) allows for a similar choice in the parameters (atlases, MRF), thus decreasing the comparison bias otherwise induced by preprocessing and parameter choice. The default value of 3 for the Mahalanobis distance, noted to be the most suitable [79] and used for comparison by García-Lorenzo et al. [110] was chosen in all experiments. As far as the MRF parameter choice is concerned in this case, two points of comparison were chosen, one taking the same H matrix defined for BaMoS (noted EMS-C in the following) and the default adaptive MRF detailed by Van Leemput et al. [163] (noted EMS-D from now on).

The second algorithm, named Lesion Segmentation Tool (LST) and developed by Schmidt et al. [90], is a toolbox of the SPM8 package and available online at `http://www.applied-statistics.de/LST_1.2.3.zip`. Among the variety of proposed method in the literature, this method has been validated both for applications in the context of Multiple Sclerosis and age-related WMH with a difference on a single threshold parameter. According to [108], the default value of 0.30 is to be chosen for MS applications (LST-MS) whereas the value of 0.25 is more appropriate when applied to age-related WMH (LST-WMH).

The third comparison point was the Lesion-TOADS (TOADS) algorithm [78] part of the MIPAV platform that belongs to the family of fuzzy methods but also corrects for IIH within its scheme is also made available at `http://www.nitrc.org/projects/toads-cruise/`.

5.5.2 Results

5.5.2.1 Brainweb simulated data

Similar measures to those assessed in Section 5.4.2 were evaluated for the external comparison. For the T1T2 modality, the various assessment measures across the different

		Assessment method							
		DSC	VD	FPR	TPR	FNR	AvDist	DE	OER
Mild	BaMoS	**51.4**	**46.4**	**41.1**	52.0	48.0	**2.7**	**40.9**	86.5
	EMS-D	2.6	4851.6	98.7	**58.3**	**41.7**	40.0	13207.9	1063.9
	EMS-C	10.7	724.3	93.9	49.1	50.9	37.4	2600.4	82.8
	LST-WML	22.8	56.1	58.8	16.0	84.0	13.4	186.8	67.2
	LST-MS	22.3	70.6	49.2	14.5	85.5	13.5	146.8	**66.7**
	TOADS	0.8	983.8	99.5	4.9	95.1	24.9	4190.6	137.3
Moderate	BaMoS	**70.2**	**22.9**	30.3	72.8	27.2	**0.8**	**173.9**	55.3
	EMS-D	20.4	582.6	88.1	**76.5**	**23.5**	16.5	13197.8	188.0
	EMS-C	44.9	91.8	65.4	67.0	33.0	15.8	2935.0	65.8
	LST-WML	62.9	30.2	23.3	53.6	46.4	3.9	693.9	**42.1**
	LST-MS	62.6	34.7	**20.5**	51.9	48.1	4.0	648.9	42.4
	TOADS	15.0	64.2	87.8	19.9	80.1	7.8	1362.3	185.0
Severe	BaMoS	**81.2**	23.2	26.1	**90.5**	**9.5**	**0.4**	**33.6**	41.7
	EMS-D	39.9	195.5	73.0	78.4	21.6	10.0	13105.7	88.7
	EMS-C	64.1	42.0	41.9	74.1	25.9	8.7	2596.9	49.6
	LST-WML	77.3	13.1	16.6	72.5	27.5	0.9	310.8	**38.8**
	LST-MS	77.1	15.8	**15.2**	71.2	28.8	0.9	286.6	38.9
	TOADS	64.6	**6.1**	33.3	62.6	37.4	2.1	838.2	60.6

Table 5.15: External comparison for the different assessment measures of the segmentation method for the T1T2 combination modality. The results are taken as the mean over all level noise and IIH at the three different lesion loads. AvDist is given in mm, DE in μL and all other assessments in %.

methods at all lesion loads are assembled in Table 5.15.

In turn the behaviour with respect to noise level for the different measures is plotted in Figure 5.29.

As in Section 5.4.3.1, the mean DSC over noise and IIH are calculated for the different lesion loads and the various algorithms across the possible modality combi-

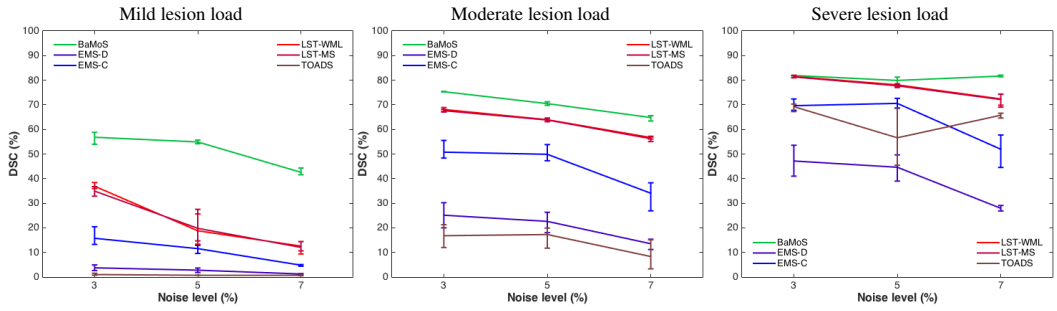


Figure 5.29: Comparison of DSC results for the automated methods with noise level variation at mild (left), moderate (middle) and severe (right) lesion load. The errorbars refer to the minimum and maximum obtained when varying the intensity inhomogeneity level.

		Modality combination			
Load	Method	T1PD	T1T2	T1T2PD	T2PD
Mild	BaMoS	11.1	51.4	47.5	18.0
	EMS-D	0.7	2.6	3.0	1.2
	EMS-C	5.7	10.7	7.1	4.9
	LST-WML	0.0	22.8	/	/
	LST-MS	0.0	22.3	/	/
	TOADS	0.8	0.8	/	/
Moderate	BaMoS	27.5	70.2	63.5	29.3
	EMS-D	7.4	20.4	24.9	14.4
	EMS-C	22.8	44.9	43.0	22.3
	LST-WML	0.0	62.9	/	/
	LST-MS	0.0	62.6	/	/
	TOADS	14.0	15.0	/	/
Severe	BaMoS	42.2	81.2	77.3	60.2
	EMS-D	13.9	39.9	46.2	34.0
	EMS-C	26.3	64.1	61.2	37.7
	LST-WML	0.0	77.3	/	/
	LST-MS	0.0	77.1	/	/
	TOADS	65.1	64.6	/	/

Table 5.16: Mean DSC (%) results over noise and IIH levels for the compared methods for various modality combinations at the three lesion loads. The slash (/) sign indicates that the combination was not possible to use for the given method.

nations and presented in Table 5.16.

As neither TOADS nor LST are able to handle the T1T2PD and T2PD combinations, the results are not provided. The visual comparison between the segmentation obtained for the automated methods (BaMoS, EMS-C, LST-MS and TOADS) is presented in Figure 5.30.

5.5.2.2 MICCAI MS data

Due to its low performance in the previous experiments, the results obtained for EMS-D are not presented in the remaining of the figures. The suggested LST-MS parameter (0.30) was used for the MS experiment and the LST-WML parameter (0.25) was used for the WMH experiment.

Using the dataset from the MICCAI MS Challenge defined in Section 5.3, comparisons between methods were performed using T1 and FLAIR images. For the assessment methods described in Table 5.3, statistical results are gathered in Figure 5.31 where each reference method (in rows), is compared against all other methods using all

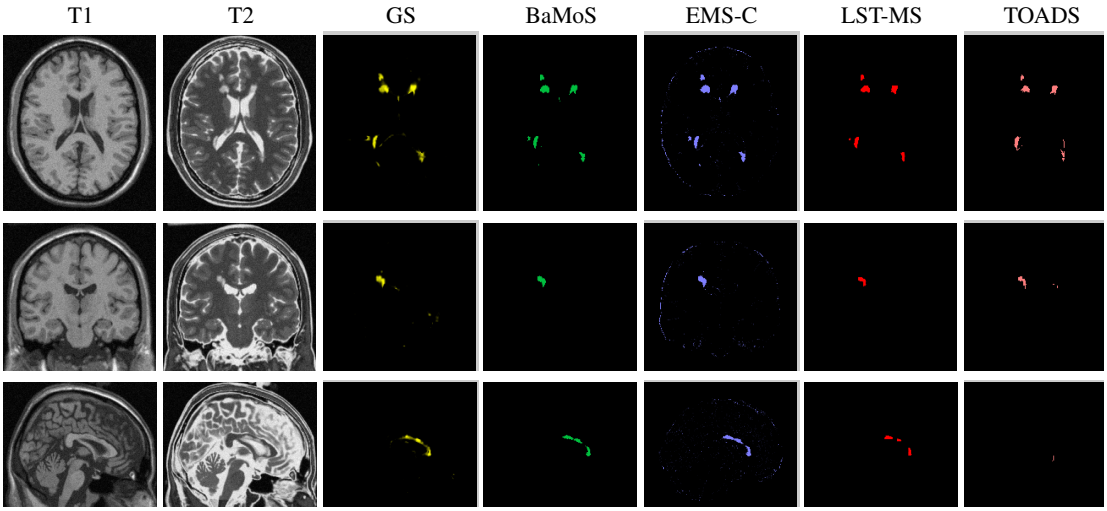


Figure 5.30: Simulated Brainweb multiple sclerosis model with severe lesion load case. Each row displays in a different orientation (axial, coronal and sagittal) from left to right the T1 image, the T2 image, the ground truth (GT) for the lesion segmentation and the corresponding results for BaMoS, EMS-C, LST-MS and TOADS.

assessment measures.

In this infographic, green corresponds to a significantly better performance, grey to a non statistically significant difference in performance and red to a significantly worse performance. For each measure, the diagonals are kept white. For this dataset, BaMoS and LST both appear to perform better than TOADS and EMS. Figure 5.32 presents an example of the obtained segmentations for the different automated methods.

5.5.2.3 T2DB data

The comparison between the automated methods is summarised in terms of statistical significance in Figure 5.33 showing that for this application, BaMoS outperformed EMS and LST and performed similarly to TOADS.

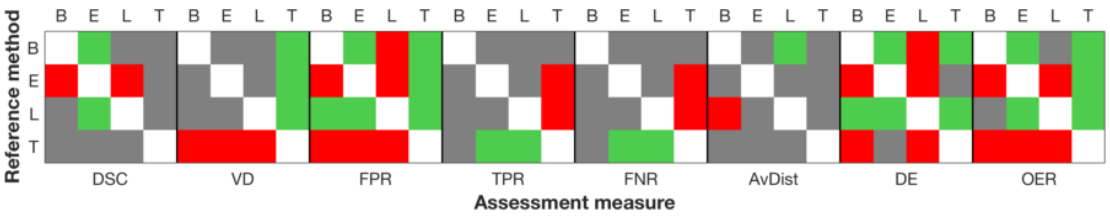


Figure 5.31: Colour-coded statistical difference significance summary for each assessment measure on the MS dataset, where each automated reference method: BaMoS (B), EMS (E), LST (L) and TOADS (T) is tested against another method (column) for a specified assessment measure. Green relates to a significantly better performance, Red to a significantly worse performance and Grey to a non statistically significant difference.

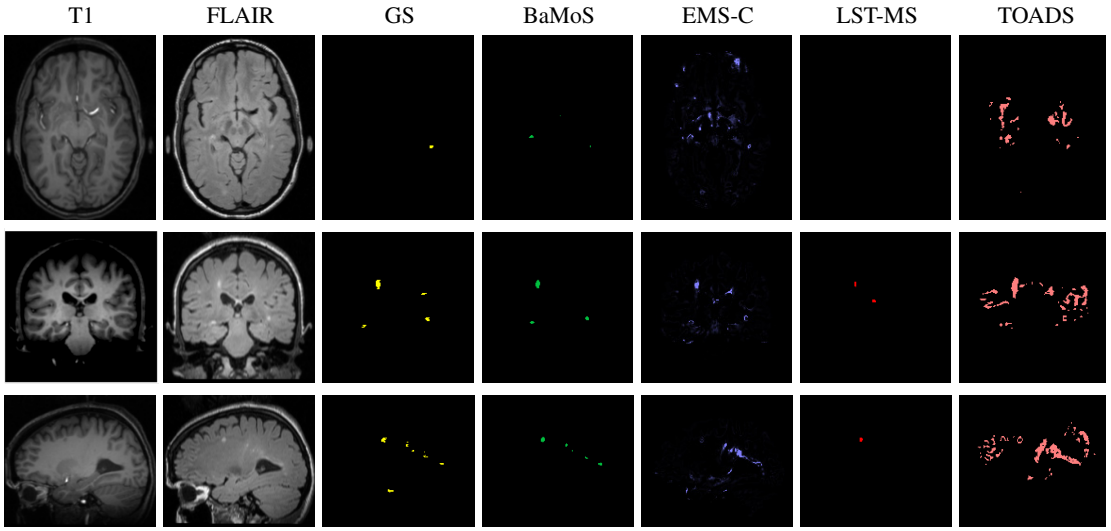


Figure 5.32: Comparison of segmentation results for an MS patient. Each row displays in a different orientation (axial, coronal and sagittal) from left to right the T1 image, the FLAIR image, the gold standard (GS) for the lesion segmentation and the corresponding results for BaMoS, EMS-C, LST-MS and TOADS.

Since the measure of global TLL has been related with cognitive decline, TLL correlations between the automatic and manual segmentations were studied using both the Pearson's R^2 correlation coefficient and the slope of the linear regression for the twenty cases. The quantitative results for both lesion segmentation assessment measures and TLL regression are presented in Table 5.17. In this table, the last line, corresponding to the TLL linear regression is performed over the 20 subjects whereas all the other assessments are summarised for subjects with a positive TLL. BaMoS slightly underestimated the lesion volume (linear coefficient of 0.88) but the correlation was high when compared to the other methods (R^2=0.96). The volumetric results are presented visually in Figure 5.34.

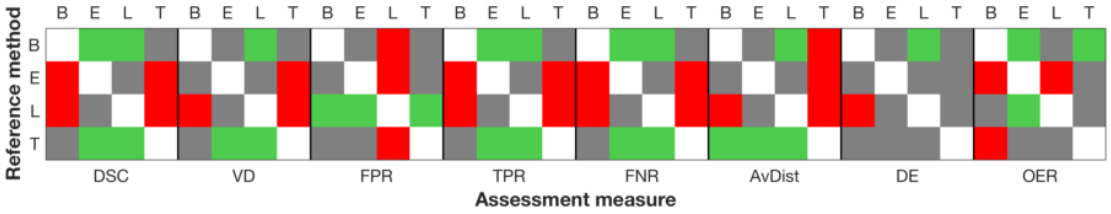


Figure 5.33: Summary of statistical differences observed between the automated methods for each assessment measure in a Green Grey Red code for the age-related WMH dataset. Each method used as a reference (row) is compared to the other three (column). Significantly better and worse performances for a specific assessment are coded in green and red respectively. No statistically significant difference is coded in grey. Diagonals stay white.

Method	BaMoS	EMS-C	LST-WML	TOADS
DSC	46.2 26.5	27.92 18.77	30.9 25.0	52.1 17.6
VD	52.0 29.0	105.6 193.8	75.7 21.4	183.2 644.0
FPR	24.4 14.2	51.4 27.5	15.1 23.0	32.1 26.1
TPR	37.7 25.8	20.9 13.9	21.7 20.1	44.5 12.9
FNR	62.3 25.8	79.1 13.9	78.3 20.1	55.5 12.9
AvDist	6.8 11.6	7.7 5.6	10.2 12.4	3.8 6.9
DE	364 215	520 330	692 392	393 254
OER	52.7 23.4	69.2 15.8	49.7 0.7	86.1 100.2
R^2 **Slope**	0.96 0.88	0.86 0.41	0.94 0.63	0.90 0.53

Table 5.17: Comparison of the different methods according to the various lesion segmentation assessment measures for the T2DB dataset of 19 subjects with positive TLL. All the measures are given in a two lines format with the mean on the first and the standard deviation on the second. AvDist is given in mm, DE in μL. The last set of lines gives the Pearson's R^2 correlation coefficient for the twenty subjects and the corresponding linear coefficient (Slope).

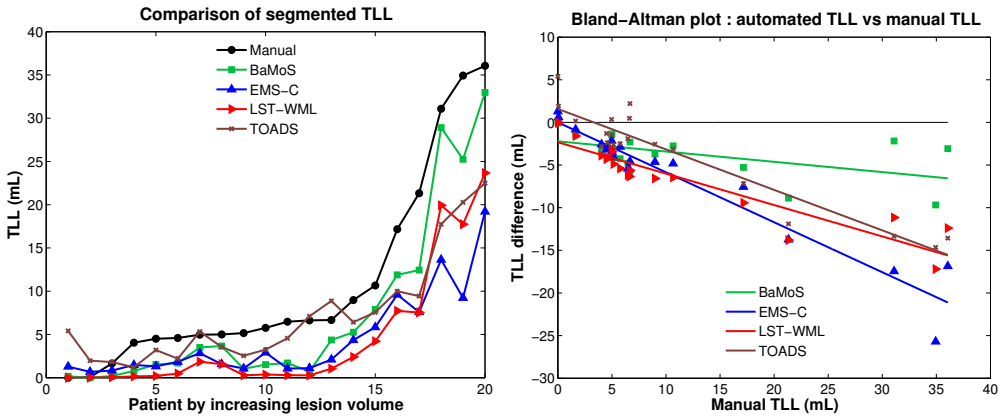


Figure 5.34: Left) Comparison of TLL per patient for the five automated methods against the manual segmentation and Right) Bland-Altman plot of the automated methods against the manual gold standard segmentation. The markers represent the twenty cases from the T2DB dataset and the line the corresponding linear fit TLL_{auto} - TLL_{manual}.

The Bland-Altman plot shows less bias for BaMoS. The plot of automated TLL per patient ordered by increasing manually segmented TLL highlights a potential problem regarding lesion overestimation in TOADS for very mild cases. The analysis of the errors observed in BaMoS with respect to the manual segmentation showed that 12% of the FN corresponded to missed lesions, the rest being related to the outline of the lesions (*i.e.* border disagreement). Among those missed lesions, 87% of this amount corresponded to lesions with a volume lower than 0.1 mL. When comparing the number of missed lesions in the automated methods, no significant difference was observed between BaMoS, EMS and TOADS, and all were able to detect significantly more lesions than LST. The periventricular region, known to be prone to partial volume effect due to resampling and shining through effects on the ventricular lining was the most prone to FN outline errors. In turn, false positives, that represent 20% of the errors occurred 65% of the time at the outline of lesions. For the erroneously detected FP lesions, 66% of the volume corresponded to lesions of less than 0.1mL. Those FP lesions were mostly located close to the ventricular lining. Thanks to the BIC constraint, the number of subclasses observed for the inlier classes was stable across the clinical dataset.

5.5.3 Discussion

With respect to the simulated data, compared to the other three families of methods, BaMoS appeared more robust to noise, especially compared to TOADS and EMS, a feature that is particularly important for milder cases of lesion load. Naturally, the number of Gaussian components needed to model the lesions decreased with an increase in the noise level. BaMoS was reasonably stable when presented with different combinations of modalities and when compared to the other automated techniques that are optimised towards specific combinations. This can be of real interest when considering clinical studies for which some imaging modalities might not be available for certain subjects. Further investigation would be needed to better understand the biological correlates between various modalities and observed signal as well as the direct impact on lesion detection and segmentation.

In the age-related WMH dataset, R^2 of the linear regression between volumes was 0.96 for BaMoS compared to 0.88 for EMS-C, 0.94 for LST-WML and 0.90

for TOADS. As mentioned in Section 5.2.1, the TLL is however insuficient in characterising the segmentation accuracy, lesion shape, localisation accuracy and overlap [50, 90, 157]. In order to better understand the origin of the segmentation errors, the eight segmentation evaluation measures defined in Table 5.3 were used to assess the automated segmentations. BaMoS obtained comparable, and many times improved results when compared to automated methods in both MS and age-related WMH contexts with a tendency to slightly underestimate the lesion volume. However, BaMoS had consistently good results when compared to the heterogeneous performance of competing methods.

An extensive comparison of the automatic results provided by BaMoS with the manual segmentation used as ground truth for the T2DB data showed a pattern in the occurrence of false negatives (FN). The corresponding voxels were usually classified as inliers of the GM or as CSF outliers. Three main types of FN can be distinguished:

Periventricular GM false negatives These FN elements are located in the periventricular regions, at locations where the WMH are similar in appearance to GM and where, due to the existence of DGM structures, the anatomical statistical atlases can further support the classification under the GM label. The presence of these FN is further explained by the fact that this area is particularly prone to partial volume effect due to image resampling.

Periventricular CSF false negatives These FN elements are located in the periventricular regions, mostly at the horns of the ventricles regions in which the corresponding CSF statistical atlas is higher than the WM one, thus enforcing the classification of the corresponding lesions as outliers spatially related to CSF. They are undetected if the CSF outliers are discarded in the lesion selection process.

Subcortical false negatives These FN elements are located near the cortical sheet and are generally undetected due to their level of hyperintensity combined with the high support of the statistical atlas for GM that do not allow to separate them from GM inliers.

Outline false negatives These FN elements are located at the border of segmented

lesions and are generally classified as WM inliers. Their level of hyper-intensity is lower compared to the local TP and in most of the cases, the corresponding T1 values are isointense compared to the WM inliers.

5.6 Consistency based validation

When focusing on the provided manual segmentations, some intensity discrepancies were observed in the classification as lesion or not especially at the lesion boundaries. The partial inclusion of the ventricles in the lesion delineation contributed for instnace to the accumulation of FN findings while as mentioned in Section 5.4.5 at the border of the lesions, slightly hyperintense FLAIR voxels that presented T1 hypointensity were in turn not considered as belonging to the lesion. Such observations led to the need for a test of the manual segmentation in terms of intensity consistency.

5.6.1 Sources of segmentation inconsistency

With this analysis of the DSC, the investigation of the origin of possible false negatives in BaMoS as presented in Section 5.5.2 showed that the undetected regions were mostly small lesions of less than 0.01 mL. Lesion segmentation accurracy was also found to be negatively correlated with the proportion lesion voxels with WM-like intensities. This observation relates strongly to the segmentation protocol defined by Filippi et al. [211] for the segmentation of MS lesions in which a conservative segmentation of lesions promoting false negatives over false positives for mildly hyperintense regions is encouraged to enable the observation of change over time.

Based on the T2DB dataset (ageing population with diabetes and/or cardiovascular risk) presenting WMH for which the data provided was already corrected for intensity inhomogeneities, the intensity distribution of the segmented WM was compared to that of the manually segmented lesions. This intensity-based assessment appears easier to perform on an age-related dataset than on an MS dataset on which iso/hypointense FLAIR regions surrounded by an hyperintense rim are considered as lesions [211].

The white matter segmentation used as reference for the distribution of normal intensities was obtained as the intersection of the resultant WM obtained for the four automated methods and further corrected to retrieve any voxel considered in the manual lesion segmentation. Since in MRI the absolute signal value is not quantitative,

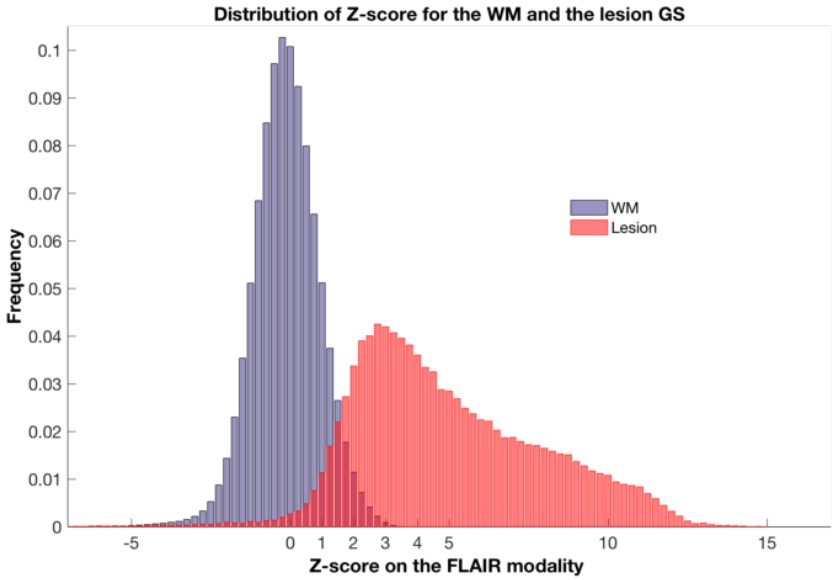


Figure 5.35: Compared Z-score distribution of manually segmented lesion and WM with respect to WM on the FLAIR modality.

Z-scores and Mahalanobis distances with respect to the mean of the WM were employed to assess the relative signal distribution. Figure 5.35 presents as an example the distribution of intensity Z-scores with respect to the WM on FLAIR modality for the manually segmented lesions and the WM mask across the dataset.

The overlap observed and the fact that the mode of the lesion segmentation is located on this overlapped region, highlights the difficulty to define consistently the limit between normal appearing white matter, dirty appearing white matter and lesions. Such uncertainty impacts directly the assessment of automated lesion segmentation methods. As an example, Figure 5.36, presents the relationship between the proportion per manually segmented lesion of voxels whose intensities falls below the threshold of 2 in Mahalanobis distance compared to the WM and the DSC for this lesion in the case of BaMoS. As expected, a negative correlation is observed, the DSC decreasing when the proportion of normal appearing voxels considered as lesion increases (R^2 0.48, Linear coefficient -0.96). Thus, assessing the automated results only based on the manual segmentation might not be sufficient to validate them and evaluating the level of systematism of such methods might be a further requirement.

5.6.2 Assessment measures - Proof of concept

Figure 5.37 illustrates such behaviours in terms of two possible assessment measures described in Section 5.2.4 namely DistQuant and PropLes. As a reminder, DistQuant

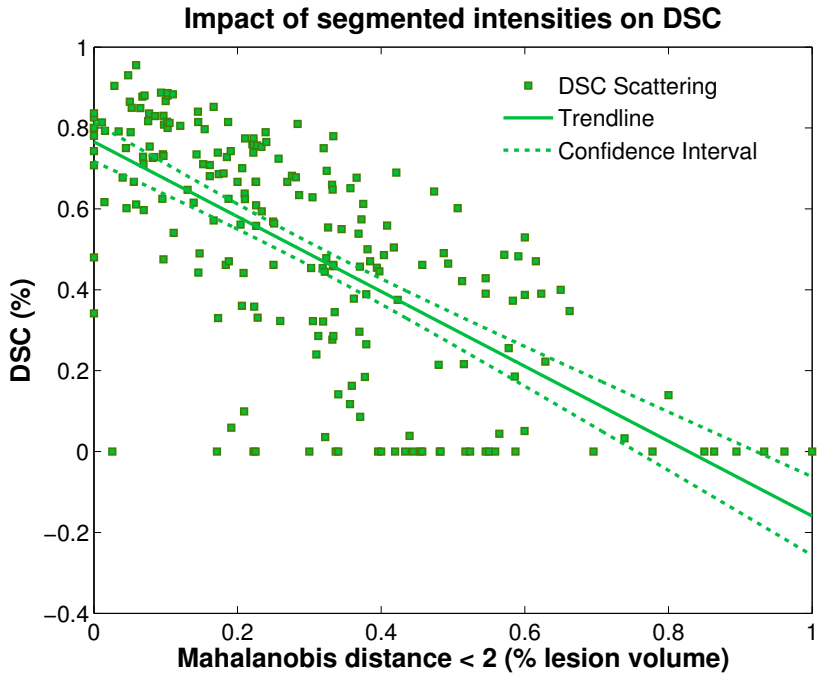


Figure 5.36: Illustration of the impact of the intensity overlap between manually segmented lesions with normal appearing white matter. The DSC per lesion (with volume >0.05 mL) is plotted (scattered points) with respect to the proportion of manually segmented voxels that present an intensity at a Mahalanobis distance inferior to 2 compared to the normal WM. The bold and dashed lines represent respectively the trendline of the correlation and the corresponding 95% confidence interval.

is the difference between the value of the first quartile of Z-scores for the lesion intensities and the third quartile of Z-scores for the normal appearing WM. Z-scores are calculated with respect to the FLAIR normal appearing white matter observations. In turn, PropLes is the proportion of voxels classified as lesion whose intensities overlap with the normal appearing WM intensity distribution. Similarly to the illustration of the intensity distributions in terms of Z-scores for the lesions and the WM, observed for the gold standard in Figure 5.35, the behaviour of the four assessed automated methods is presented in Figure 5.38. The distribution of lesion intensities is very similar for BaMoS and TOADS with a mode for the lesion intensities close to the extremal value for the normal WM. As expected from Figure 5.37, the overlap between NAWM and lesion is higher for TOADS. The shift of the mode for lesion intensities to higher values of the Z-score in the case of LST, further highlights the risk of a conservative lesion segmentation.

A high consistency of the algorithm behaviour across the dataset would correspond to a low variance for the intensity assessments. Across all methods, DistQuant

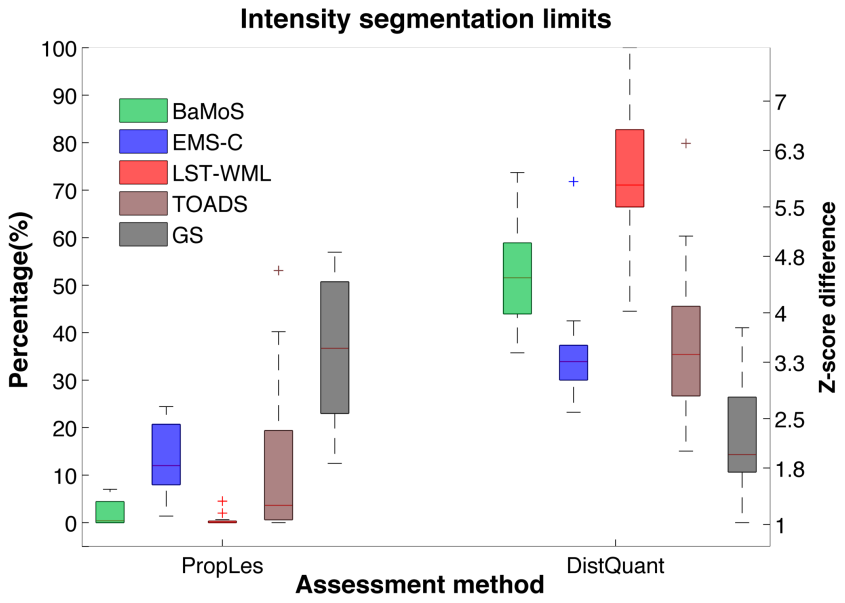


Figure 5.37: Comparison of consistency in terms of lesion standardised intensities for the four automated methods and the ground truth in terms of PropLes(left axis) and DistQuant (right axis).

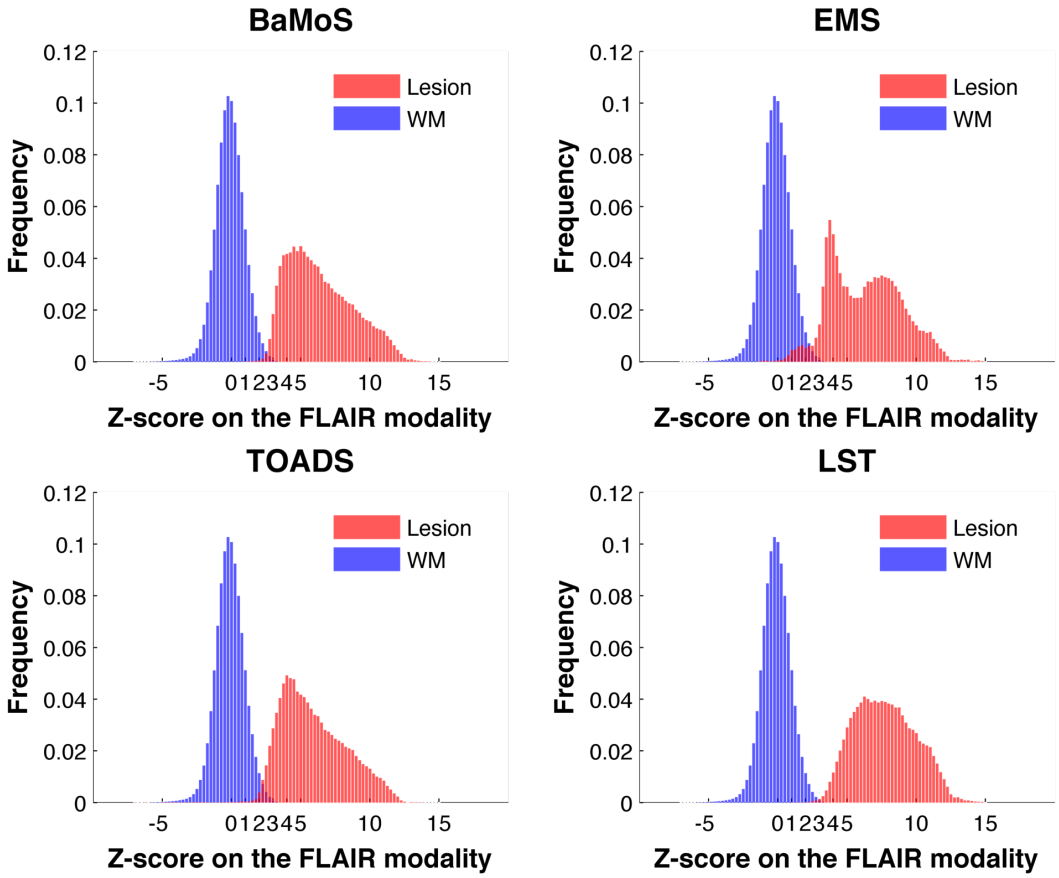


Figure 5.38: Compared Z-score distribution of automated segmented lesions with respect to WM on the FLAIR modality for BaMoS, EMS, TOADS and LST.

was observed to produce the least variations through the dataset. A low proportion overlap between intensities of normal appearing white matter and segmented lesions as was observed for LST-WML and BaMoS with a low variance satisfies the problem of consistency in terms of definition of a lesion compared to white matter. However, associated with a large DistQuant value, this can also be the sign of a tendency to classify the lesion outline as GM or to a conservative definition of lesions. The risk of including CSF voxels in the final TOADS segmentation expresses itself in the larger proportion of lesion intensities overlapping with NAWM intensities. The large overlap observed between NAWM and lesion intensities for the gold standard segmentation as observed in Figure 5.35 is represented with these consistency measures with the high PropLes and low DistQuant.

Thus, the measures developed may reflect consistency across a dataset through the estimation of their variation while their mean values may be representative of the behaviour of a given algorithm thus allowing for more principled comparisons.

5.7 Discussion

5.7.1 Synthetic data

When compared to the validation on clinical data, the main advantage of using synthetic images is the availability of a ground truth. Additionally the possibility of modelling different conditions of image quality enables to test the robustness of a given algorithm to these effects. The Brainweb data was used here to test the algorithmic stability to different imaging modality combinations, different degrees of image quality with varying noise level and intensity inhomogeneity. However, the validation using the Brainweb dataset is limited by multiple factors as previously underlined by García-Lorenzo et al. [110]. As only one phantom is available, no statistical analysis is possible. Also, synthetic MR simulated images cannot be considered truly realistic. Moreover the range of lesion load is limited compared to the amount that can be found in clinical cases. Furthermore, although both MS and age-related WMH are evolutionary processes, there is no longitudinal component in the lesion simulation. Such a longitudinal simulator based on lesion transfer on clinical data will be developed in Section 7.3.1.

5.7.2 Local disagreements

Consistently across experiments, systematic areas appeared to be prone to misclassifications in terms of lesion segmentation. With respect to the preprocessing choice, the impact of modality choice and the acquisition protocol, the infratentorial regions appeared indeed commonly subject to disagreement.

Different reasons may be presented to explain this local finding. First, the infratentorial region is known to present a different contrast in the FLAIR acquisition compared to the cerebral white matter. Therefore, a change in normalisation due to mask variation may first lead to segmentation differences as in Section 5.4.1. Such slight hyperintensities may also be initially considered as susceptible outliers when using first the FLAIR image instead of the T2 modality when all three are available as described in Section 5.4.3. Such observations led for instance Samaille et al. [95] to adopt specific size rules to account for potential false positives in this area and Datta et al. [99] to separate the analysis of the cerebellum from the rest of the image.

Second, critical flow and pulsation artefacts combined with shine through effects may be present along the Sylvian aqueduct and the fourth ventricle. The false positive correction supposed to account for these findings depends however heavily on the statistical priors, healthy tissue segmentation and distance to mask borders, which may lead to different decisions on the considered lesion elements. Such artefacts, frequently present in 2D acquisitions of FLAIR images, are reported to be largely reduced in 3D acquired images [161], observation confirmed in the experiment on the POPPY dataset (cf Section 5.4.4).

In the various situations, the ventricular lining appeared also prone to segmentation disagreement. As this region has been highlighted as of special interest when dichotomizing the clinical impact of white matter hyperintensities [18], further work would be needed to avoid confusion between deep gray matter and WMH. In order to consistently assess the location of areas of greater uncertainty, one can develop systematic location schemes as will be shown in Chapter 6.

5.7.3 Generalisation and segmentation uncertainty

In this segmentation framework, elements of the model that correspond to WMH are extracted automatically after the model optimisation, providing the final lesion segmen-

tation described in Section 5.1.

As the model selection framework is independent of the observed pathology, BaMoS has a large clinical flexibility to different modalities and pathological contexts. In fact, the final data model is independent from the lesion definition. Segmentation rules can thus be adapted to the variability in lesion definition contrarily to other models in which heuristic rules are included in the cost function optimisation. This flexibility allows for more variation in the clinical description of lesions and further emphasises the generalisation potential of BaMoS. Thanks to the generic nature of BaMoS, associations of different types of outliers such as white matter hyperintensity and iron deposition could be for instance investigated, since iron deposition is known to be associated with cognitive decline in ageing [194] and observed in the course of MS [212]. Although other than purely biological causes such as partial volume or bias field may cause model components to split, the separation of lesion elements into different classes according to their intensity opens the door to a deeper understanding of the underlying lesion's pathophysiology. Besides, these levels of intensity may be of further interest when studying the evolution of the WMH longitudinally since in the elderly it has been shown that such lesions evolve from pre-existing WM damages [213].

Chapter 6

WMH spatial distribution

Although global WMH lesion burden is reported to be associated with clinical cognitive outcome and various risk factors, further description of WMH appears to be able to refine clinical correlations and assessments. Longitudinally, progression of WMH has been observed to occur from preexisting lesions [214]. Investigating the progression of WMH lesions according to their initial degree of confluency (focal, early confluent, confluent), a pictorial characteristic often used in visual rating scales, Enzinger et al. [215] showed differences in the progression patterns of the different lesion groups with no progression for the focal lesions and a strong accumulation in the case of the confluent damage. Such distinction is mainly related to the underlying severity of the pathology [19]. Punctuate lesions correspond usually to a mild tissue damage confined to an area surrounding a dilated periventricular space whereas confluent regions of hyperintensities are associated to diffuse areas of incomplete parenchymal destruction with myelin and axonal loss and correspond to a continuum of tissue damage [16]. As underlined by Enzinger et al. [215], using lesion appearance to select a population can considerably affect the design of clinical trials with respect to the required sample sizes.

Other than appearance, lesion spatial distribution has been studied both in cross-sectional and longitudinal investigations on the clinical correlates of WMH. For instance, progression of WMH in the parietal lobe has been reported to be a predictor of Alzheimer's Disease [216]. The variability in image acquisition and definition of region of interest have been highlighted as potential causes for the mixed and controversial findings in the relationships between lesion location and clinical outcome [35]. Lesion spatial distribution is nonetheless prominent in the description of visual rat-

ing scales, notably when distinguishing between periventricular (PV-) and deep white matter hyperintensities (DWMH) in their assessment. The Fazekas scale [217] and the Scheltens scale [218], both widely used clinically, include this separation in their description of lesions.

As mentioned in Section 1.1.2.2, such distinction has been partly substantiated by histopathological findings [18, 219] differentiating the ischaemic pathway leading to DWMH from the the loosening of white matter in the most periventricular lesions. According to the study performed by Haller et al. [20], level of demyelination assessed via observation of FLAIR images tended to be overestimated in PV regions compared to DWM due to the higher interstitial water content in periventricular regions. Comparatively, Murray et al. [44] noticed that the oligodendrocyte population was less affected in DWMH than in PVWMH and that weaker hyperintense lesions did not see the oligodendrocytes affected. This observation was used to infer an order in the degradation of the WM, the myelin being affected before reaching the core of the oligodendrocytes.

Functionally, this distinction has been assessed in various studies emphasising the effect of periventricular lesions on executive functioning [35] and of deep white matter lesions on mood and depression in cross-sectional [33, 220] and longitudinal studies [42]. The segregation and definition between periventricular and deep white matter hyperintensities is however known to suffer from various drawbacks and has been challenged. De Carli et al. [36] underlined the noticeable correlation between DWMH, PVWMH and total lesion load while Barkhof et al. [221] explained a reported 80% of PVWMH as resulting from characterisation issues. If the main definition of periventricular lesion relies on the property of continuity of the lesion with the ventricular surface, other criteria such as maximal distance to the ventricles, shape of the lesion or maximal extent have also been proposed to account for the common but delicate case of coalescing lesions [18]. The use of absolute value for the maximal distance when qualitatively scaling or automating the separation [222, 223] is however problematic in an ageing population due to the concomitant brain atrophy and ventricular expansion and the variation between subjects and is not supported by any biological meaning. As an alternative to this dichotomisation, with the development of automated segmentation methods, lesion probability maps and the voxelwise relationships with clinical outcome have also been studied [30]. These methods suffer however from the

low voxelwise probabilities of lesion occurrence [27].

In this chapter, a systematic patient-specific coordinate frame is developed with respect to both the distance to the ventricular surface and the lobar position. Comparatively, the closest existing location scheme for WMH is the one developed by Van der Lijn et al. [224] in which the brain is regionally separated into an anterior, posterior and central region and fifteen layers of absolute distance to the ventricular surface are drawn. After describing the process of projecting 3D lesion information to a standardised coordinate frame, the relevance of this projection using two examples is explored. First, the spatial distribution of lesions is studied in a population of pairs of monozygotic twin subjects. Then, the developed scheme is applied to the deconstruction of visual rating scales for WMH, leading to the creation of a training tool for radiologists.

6.1 Methods

6.1.1 Relative ventricular distance

The aim of the relative ventricular distance is to define a WM radial coordinate system between the ventricular CSF and the cortical GM.

Similarly to cortical thickness, the WM coordinate frame must be smooth and allow for bijective correspondences with trajectories normal to the boundary surfaces on both sides. As proposed by Jones et al. [225], series of equipotential nested surfaces that fulfil these conditions can be described by the Laplace's equation. According to it, a field ψ enclosed between boundaries (here the ventricular surface and the cortical sheet surface) with different boundary conditions for the potential satisfies

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = 0.$$

Field lines relating the two surfaces are obtained by integrating the normalised gradient of the field between the two surfaces. The Laplace equation can be solved numerically using iterative solutions such as the Jacobi method, referred to by Jones et al. [225]. At any point (x,y,z) of the volume considered, the second order Taylor expansion of the field for all neighbours is given by

$$\psi(x+\Delta x,y,z) = \psi(x,y,z) + \Delta x\frac{\partial\psi(x,y,z)}{\partial x} + \frac{\Delta x^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial x^2} + O(\Delta x^3)$$

$$\psi(x-\Delta x,y,z) = \psi(x,y,z) - \Delta x\frac{\partial\psi(x,y,z)}{\partial x} + \frac{\Delta x^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial x^2} + O(\Delta x^3)$$
$$\psi(x,y+\Delta y,z) = \psi(x,y,z) + \Delta y\frac{\partial\psi(x,y,z)}{\partial y} + \frac{\Delta y^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial y^2} + O(\Delta y^3)$$
$$\psi(x,y-\Delta y,z) = \psi(x,y,z) - \Delta y\frac{\partial\psi(x,y,z)}{\partial y} + \frac{\Delta y^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial y^2} + O(\Delta y^3)$$
$$\psi(x,y,z+\Delta z,) = \psi(x,y,z) + \Delta z\frac{\partial\psi(x,y,z)}{\partial z} + \frac{\Delta z^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial z^2} + O(\Delta z^3)$$
$$\psi(x,y,z-\Delta z) = \psi(x,y,z) - \Delta z\frac{\partial\psi(x,y,z)}{\partial z} + \frac{\Delta z^2}{2}\frac{\partial^2\psi(x,y,z)}{\partial z^2} + O(\Delta z^3).$$

With a first order approximation, the Laplace equation becomes

$$\begin{aligned}
&\frac{\psi(x+\Delta x,y,z)-2\psi(x,y,z)+\psi(x-\Delta x,y,z)}{\Delta x^2}+\\
&\frac{\psi(x,y+\Delta y,z)-2\psi(x,y,z)+\psi(x,y-\Delta y,z)}{\Delta y^2}+\\
&\frac{\psi(x,y,z+\Delta z)-2\psi(x,y,z)+\psi(x,y,z-\Delta z)}{\Delta z^2} = 0\\
&\frac{1}{2}\left(\frac{\psi(x+\Delta x,y,z)+\psi(x-\Delta x,y,z)}{\Delta x^2}+\right.\\
&\frac{\psi(x,y+\Delta y,z)+\psi(x,y-\Delta y,z)}{\Delta y^2}+\\
&\left.\frac{\psi(x,y,z+\Delta z)+\psi(x,y,z-\Delta z)}{\Delta z^2}\right) = \frac{\psi(x,y,z)}{\Delta x^2}+\frac{\psi(x,y,z)}{\Delta y^2}+\frac{\psi(x,y,z)}{\Delta z^2}.
\end{aligned}$$

The Jacobi solution consists in progressively updating at iteration $t+1$

$$\begin{aligned}
\psi^{(t+1)}(x,y,z) =& \frac{1}{2}\left(\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}+\frac{1}{\Delta z^2}\right)^{-1}\\
&\cdot\left(\frac{\psi^{(t)}(x+\Delta x,y,z)+\psi^{(t)}(x-\Delta x,y,z)}{\Delta x^2}\right.\\
&+\frac{\psi^{(t)}(x,y+\Delta y,z)+\psi^{(t)}(x,y-\Delta y,z)}{\Delta y^2}\\
&\left.+\frac{\psi^{(t)}(x,y,z+\Delta z)+\psi^{(t)}(x,y,z-\Delta z)}{\Delta z^2}\right).
\end{aligned}$$

This iterative process is conducted until the relative change in energy $\frac{\varepsilon_{t+1}-\varepsilon_t}{\varepsilon_t}$ reaches a threshold ε, the field energy being at iteration t over the N voxels of image I

$$\varepsilon_t = \sum_I \left[\left(\frac{\Delta\psi}{\Delta x}\right)^2+\left(\frac{\Delta\psi}{\Delta y}\right)^2+\left(\frac{\Delta\psi}{\Delta z}\right)^2\right]^{1/2}.$$

As mentioned by Yezzi et al. [226], the values of the harmonic solution on the volume do not increase proportionally to the arclength following its gradient flow lines. Therefore, there is no defined level set value of the solution that will cut the trajectories at a

certain relative length. As proposed by Yezzi et al. [226], a normalised length function can be defined in order to allow level sets to be drawn directly. Introducing $\mathbf{N} = \frac{\nabla \psi}{\| \nabla \psi \|}$ the normalised tangent field to the trajectories flowing from the inner (ventricular) surface to the outer (cortical) surface, the length along the path coming from the inner surface L_I verifies $\nabla L_I \cdot \mathbf{N} = 1$ while the length along the path coming from the outer surface L_O satisfies $-\nabla L_O \cdot \mathbf{N} = 1$, and the length being equal to 0 at their respective starting point. The normalised distance from the inner surface at any point P of the region R is then equal to $\frac{L_I(P)}{L_I(P)+L_O(P)}$. To obtain the distance maps in the case of images, the backward and forward differences are introduced denoting h_t the voxel size in the t direction and i, j, k the indices in the x, y and z direction respectively so that

$$
\begin{array}{lll}
D_x^- L = \frac{L(i,j,k) - L(i-1,j,k)}{h_x} & D_y^- L = \frac{L(i,j,k) - L(i,j-1,k)}{h_y} & D_z^- L = \frac{L(i,j,k) - L(i,j,k-1)}{h_z} \\
D_x^+ L = \frac{L(i+1,j,k) - L(i,j,k)}{h_x} & D_y^+ L = \frac{L(i,j+1,k) - L(i,j,k)}{h_y} & D_z^+ L = \frac{L(i,j,k+1) - L(i,j,k)}{h_z}
\end{array}
$$

Since the gradient can be defined either using backward or forward differences, the solution is obtained in the direction of the tangent field, thus determining the differentiation scheme to adopt. Therefore, the equation becomes

$$
\begin{aligned}
1 =& N_x(i,j,k) \begin{cases} D_x^-(i,j,k) \text{ if } N_x(i,j,k) > 0 \\ D_x^+(i,j,k) \text{ otherwise} \end{cases} \\
&+ N_y(i,j,k) \begin{cases} D_y^-(i,j,k) \text{ if } N_y(i,j,k) > 0 \\ D_y^+(i,j,k) \text{ otherwise} \end{cases} \\
&+ N_z(i,j,k) \begin{cases} D_z^-(i,j,k) \text{ if } N_z(i,j,k) > 0 \\ D_z^+(i,j,k) \text{ otherwise} \end{cases} \\
=& \frac{N_x(i,j,k)}{h_x} \begin{cases} L(i,j,k) - L(i-1,j,k) \text{ if } N_x(i,j,k) > 0 \\ L(i+1,j,k) - L(i,j,k) \text{ otherwise} \end{cases} \\
&+ \frac{N_y(i,j,k)}{h_y} \begin{cases} L(i,j,k) - L(i,j-1,k) \text{ if } N_y(i,j,k) > 0 \\ L(i,j,k+1) - L(i,j,k) \text{ otherwise} \end{cases} \\
&+ \frac{N_z(i,j,k)}{h_z} \begin{cases} L(i,j,k) - L(i,j,k-1) \text{ if } N_z(i,j,k) > 0 \\ L(i,j,k+1) - L(i,j,k) \text{ otherwise} \end{cases}
\end{aligned}
$$

The solutions for $L_I(i,j,k)$ and $L_O(i,j,k)$ are then:

$$L_I(i,j,k) = \frac{1 + \frac{|N_x|}{h_x} L_I(i \mp 1, j, k) + \frac{|N_y|}{h_y} L_I(i, j \mp 1, k) + \frac{|N_z|}{h_z} L_I(i, j, k \mp 1)}{\frac{|N_x|}{h_x} + \frac{|N_y|}{h_y} + \frac{|N_z|}{h_z}}$$

$$L_O(i,j,k) = \frac{1 + \frac{|N_x|}{h_x} L_O(i \pm 1, j, k) + \frac{|N_y|}{h_y} L_O(i, j \pm 1, k) + \frac{|N_z|}{h_z} L_O(i, j, k \pm 1)}{\frac{|N_x|}{h_x} + \frac{|N_y|}{h_y} + \frac{|N_z|}{h_z}}$$

where N is taken in (i,j,k),

$$i \pm 1 = \begin{cases} i+1 & \text{if } N_x > 0 \\ i-1 & \text{otherwise} \end{cases} \quad j \pm 1 = \begin{cases} j+1 & \text{if } N_y > 0 \\ j-1 & \text{otherwise} \end{cases} \quad k \pm 1 = \begin{cases} k+1 & \text{if } N_z > 0 \\ k-1 & \text{otherwise} \end{cases}$$

and

$$i \mp 1 = \begin{cases} i-1 & \text{if } N_x > 0 \\ i+1 & \text{otherwise} \end{cases} \quad j \mp 1 = \begin{cases} j-1 & \text{if } N_y > 0 \\ j+1 & \text{otherwise} \end{cases} \quad k \mp 1 = \begin{cases} k-1 & \text{if } N_z > 0 \\ k+1 & \text{otherwise} \end{cases}$$

The solutions are implemented using a fast-marching algorithm. Indeed, since $\mathbf{N}$ is determined and the solution at point (i,j,k) depends on the value of three neighbours, fixing the known values can allow the solution to be obtained with only one pass through the volume elements. Once the normalised distances are obtained, a discretisation in layers is possible. Based on the indications of Kim et al. [18], four equally spaced layers are separated by default. The layers numbering increases with the distance to the ventricular surface.

6.1.2 Lobar separation

In order to inform the anatomical location of the lesions, frontal, parietal, occipital and temporal regions are defined for the right and the left hemispheres. Basal ganglia and infratentorial regions (denoted BGIT) are combined. The regional demarcation originates from the brain parcellation obtained through the GIF pipeline [198]. Multiple cortical labels are aggregated into four lobes per hemisphere and BGIT. Euclidean distance maps are then calculated from these cortical lobar regions and WM voxels are classified into corresponding regions based on a minimal distance rule. Infratentorial regions are simply derived from the parcellation while the basal ganglia segmentation is dilated in order to enclose the WM present between the putamen and the caudate or thalamus. An alternative to build the lobar separation could have been to propagate the

labels along the trajectories derived from the Laplace equation or to propagate existing atlases.

6.1.3 Display

From the lobar division and the layer discretisation, a total number of 36 brain regions is formed (9 brain regions $\times$ 4 layers). Scalar characteristics relative to the lesions in each region can then be summarised in a single plot through the use of relevant color scales. In such a plot, represented as a bullseye of concentric layers, the layer number increases radially while the lobar regions are positioned to reflect their neighbourhood relationships. The first few letters for each lobe are mentioned, using for instance FRONT to refer to the frontal lobe. Right and left are indicated on the plot. Relevant scalar characteristics are, among others, the proportion of regional volume occupied by lesions (lesion frequency) or the proportion of the total lesion load located in a specific region (lesion distribution). An example of lesion segmentation, lobar separation and layer discretisation is displayed in Figure 6.1 for the three orientations and is followed by the representation of lesion frequency (values increasing from green to red) and lesion distribution (values increasing from blue to pink).

The two graphical displays are complementary and can be interpreted as follows in this case: the lesion frequency plot expresses that the parietal periventricular zone are largely affected by lesions while the lesion distribution plot relates that a large proportion of the overall lesion volume is actually located in the intermediate zone of the parietal region.

6.2 WMH spatial distribution in twin population

The standardised description of lesion spatial distribution and the corresponding regional lesion descriptors were applied to the study of a twin population in order to illustrate the relevance of such a scheme in the assessment of potential similarities of WMH patterns within twin pairs. After contextualising brain studies with respect to twin population in Section 6.2.1, the data and performed experiments are detailed in Section 6.2.2 before presenting the obtained results (Section 6.2.3).

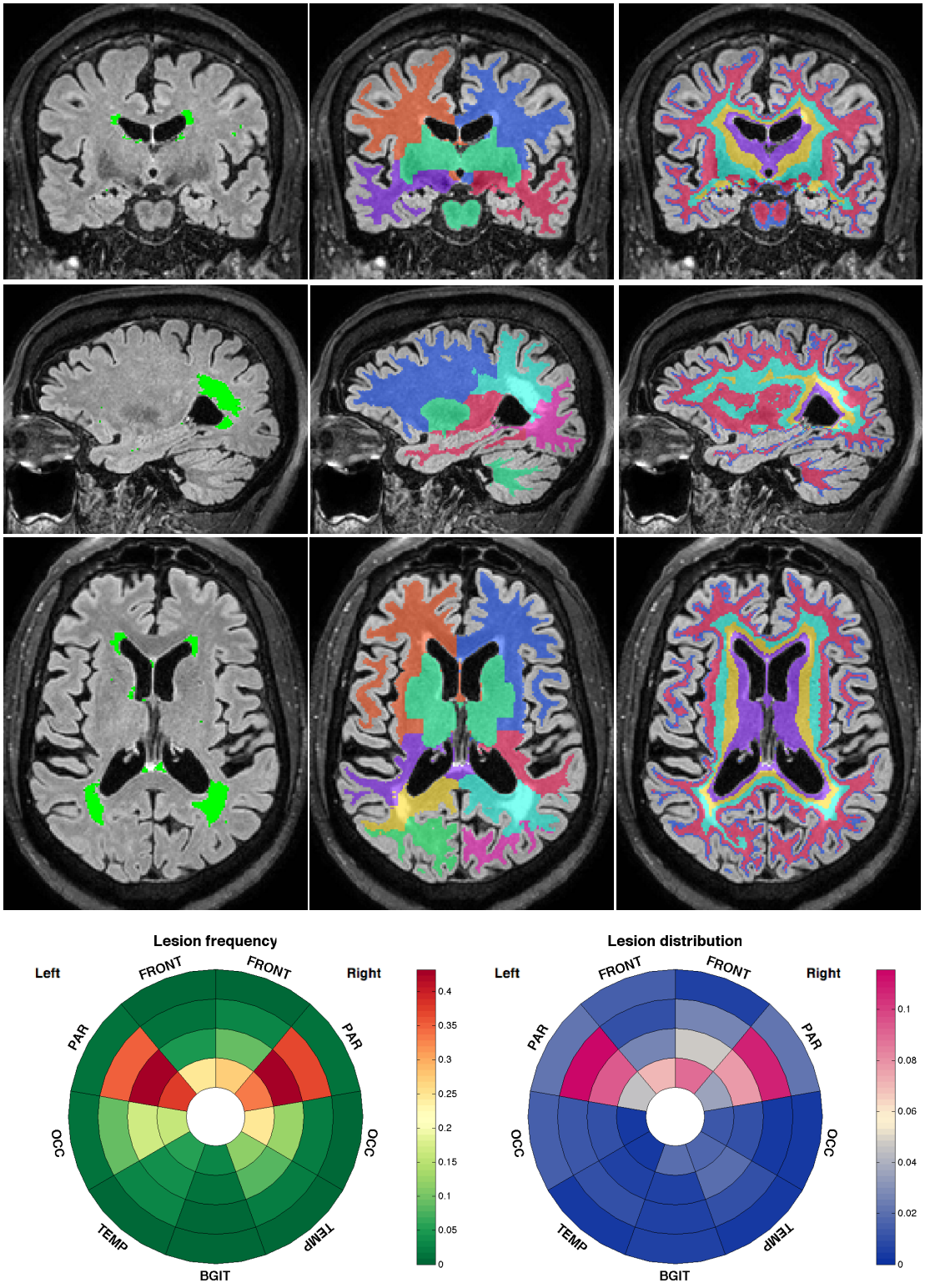


Figure 6.1: Example of the construction of regional lesion features in lobes and layers and display of the lesion frequency and lesion distribution on planar bullseyes.

6.2.1 Clinical context

The assessment of heritability in the context of pathology is especially important to better understand the interactions between environmental conditions and genetics. Studies on populations of twins are conducted in order to assess the heritability of phenotypical traits, that is the genetic component of the population variance. In turn, measures of environmental factors can be derived from the differences observed within twin pairs. In such studies, high heritability has been observed in brain volumes, cortical thickness and cortical surface [227] while the gyral structures [228] or the white matter tracts in children [229] appear less correlated. With respect to dynamic changes observed with ageing, a high heritability is reflected in total brain volumes [230, 231] and in the corpus callosum shape [230]. Besides, strong genetic factors appear to partially govern cerebrovascular pathology with for instance genetic predisposition to ischaemic strokes [232]. Additionally, in the case of small vessel disease, correlations of total lesion volumes appear to be significantly greater in monozygotic than dizygotic pairs of twins [233]. However to date, there has been no investigation on potential genetic based similarities in the pattern of WMH spatial distribution. Furthermore, since the strength of heritability varies across the brain, it would be of interest to see if such differences also exist in the case of WMH.

6.2.2 Data and Experiments

For this application, a data subset of the PreclinAD study with 43 pairs of monozygotic twins recruited from the Netherlands Twin Registry was analysed. This study aims at describing the biomarkers for amyloid pathology and cognitive decline in a cognitively healthy elderly population. Following the study protocol, both members of a twin pair undergo on the same day a battery of neuropsychological and cognitive tests, an MR session and biosamples draws. During the MR sequence on a Philips 3T Achieva scanner, T1-weighted and FLAIR sequences are among the acquired pulse sequences with the following characteristics: 1) T1-weighted 3D Fast Field Echo; TR = 7.9 ms, TE = 4.5 ms; voxel size $1.0 \times 1.0 \times 1.0$ mm^3 and 2) Fluid attenuated Inversion Recovery : TR=4800 ms, TE=279 ms, TI= 1650 ms; voxel size $0.56 \times 1.04 \times 1.04$ mm^3.

BaMoS was applied to the T1 and FLAIR images, registering the FLAIR image affinely to the T1 space and using anatomical atlases and mask from the label fusion

GIF pipeline [198]. The WMH were segmented and the local regions constituting the bullseye plot (4 layers and 9 lobar regions) were built. In order to expose the relevance of the local information, correlations between vectors of local characteristics were calculated while absolute differences in log and raw global volumes were also derived within twin pairs. For each twin pair, a random pairing with one of the individuals of the remaining pairs was performed resulting thus in 42 additional correlation values (or difference values) for both twins in the studied pair. The Pearson correlation coefficient R between the local characteristics of subject X and subject Y with N the number of characteristics is expressed as

$$R = \frac{1/N \sum_{n=1}^{N} (x_n - \mu_x)(y_n - \mu_y)}{\sigma_x \sigma_y},$$

where μ_x (resp. μ_y) is the mean of the characteristics for subject X (resp. Y) and σ_x (resp. σ_y) the corresponding standard deviation. Sorting the correlations from highest to lowest and the differences from lowest to highest, the rank corresponding to the twin pairing was determined. 1000 repeats for the random pairing were performed. The statistics for each tested characteristics were evaluated on the mean over these 1000 random draws. The rest of the analysis, performed in Amsterdam is still undergoing at the time of writing.

6.2.3 Results

Over the whole set of pairs of twins among which 67.4% were women, the volume of detected lesions was of a median of 1.99 mL (IQR=[0.76 5.58]) and the average given Fazekas score was of 1.08 (SD = 0.86). The mean age of the pairs was of 68.3 years (SD = 8.14) and the total intracranial volume as obtained applying [198] was of 1567 mL (SD = 215). A Kendall tau correlation of 0.64 was found between Fazekas scores and the segmented lesion volumes. For a twin pair, an example of the FLAIR images, resultant lesion segmentation and overall lesion frequency and lesion distribution plots is presented in Figure 6.2 to expose the strong similarities that may occur in terms of WMH distribution in the brain of twin subjects.

As a supplementary illustration, the comparison within pairs of lesion frequency (proportion of lesion occupying a regional volume) and lesion distribution (proportion of the total lesion load in the considered region) is displayed in Figure 6.3 for all pairs

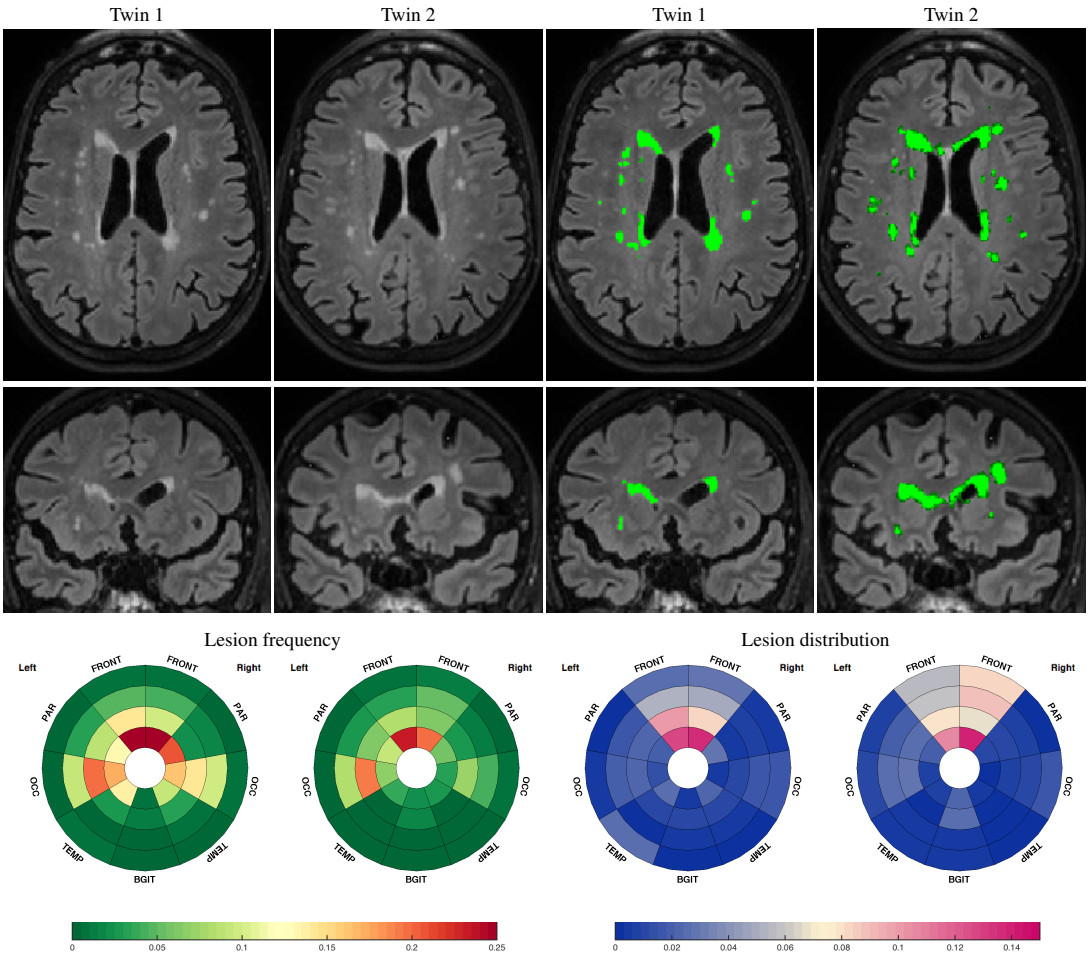


Figure 6.2: Example of similarities in WMH distribution in a twin pair. Images in axial (top row) and coronal (middle row) view are presented for each twin with the corresponding lesion segmentation. The bullseye representation of lesion frequency and lesion distribution for each twin are displayed on the bottom row.

of twins. Regarding the ranking of the within pair correlations compared to random pairing, Table 6.1 gathers the results for the local volumes, lesion frequency, lesion distribution and local log-volumes as well as global volumes and log-volumes knowing that a random pairing would give on average a ranking of 21.5 (number of pairs divided by 2). The assessment of pairwise difference using Wilcoxon paired test showed significant differences between local characteristics and global volumetric random results, the local characteristics leading to lower ranking (p-value <0.00001). The ranking obtained through local features were significantly lower than those obtained with the global volumetric measurements but in all cases the obtained ranking was lower than what random ranking would provide.

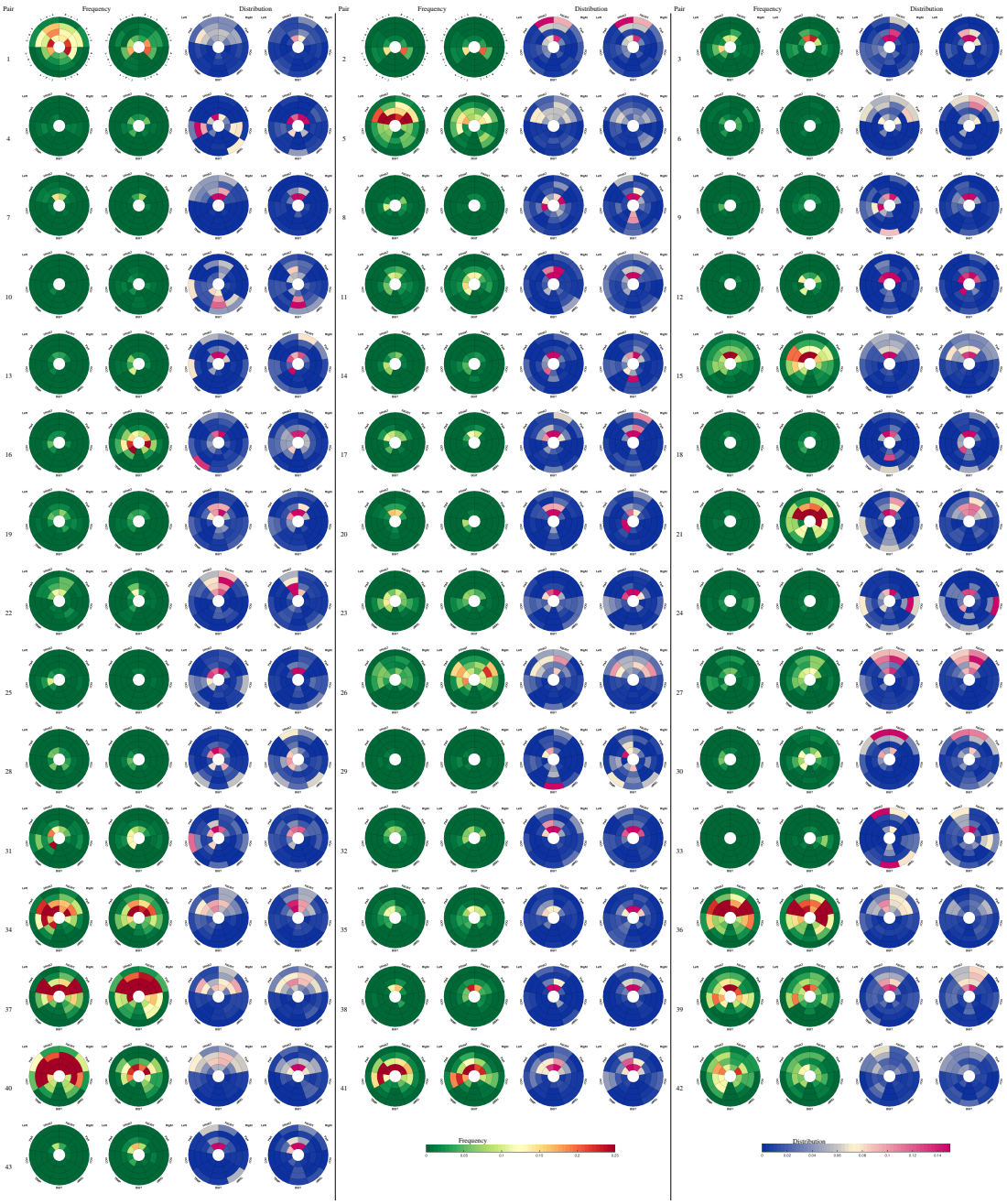


Figure 6.3: Comparison within twin pairs of lesion frequency and lesion distribution. For each pair, the lesion frequency are given in the first two plots and followed by the lesion distribution.

	Local Volumes	Local Frequencies	Local Distributions	Local Log-Volumes	Global Volume	Global Log-volume
Mean	9.90	11.12	9.88	6.40	13.63	13.60
SD	9.65	11.37	9.62	7.78	10.66	10.71
Range	[1 36.49]	[1 41.48]	[1 36.46]	[1 38]	[1.48 43]	[1 43]
Median	5.7	6.2	5.8	3.0	10.5	10.5
IQR	[2.5 15.97]	[2.503 16.821]	[2.49 15.98]	[2 7]	[5.52 18.48]	[5 19]

Table 6.1: Statistics on the ranking of the correlations or volume difference when considering twin pairs to random pairs of subjects.

6.2.4 Discussion

In this initial application of the systematic lesion spatial description scheme, patterns of similarity appeared within twin pairs as shown in Figures 6.2 and 6.3. Comparing random pairing of twins to the true pairing showed that global volumetric measurements of WMH were indeed related as reported by Carmelli et al. [233]. The proof of concept experiment based on the correlation of local information on lesions allowed twin pair specific information to be more readily available. This was shown by the lower ranking systematically obtained when using local information compared to global measures. It must however be underlined, that in this ranking experiment, neither age nor TIV was taken into account. In the case of WMH spatial distribution, in which some regions are more informative than others, considering all characteristics equally as done with the Pearson correlation coefficient, may however not be the optimal solution. Greater twin pairs concordance could certainly be obtained when using weighted correlation relationships. To some extent the log-transformation of the volumes reflects the effects of a reweighing and showed here the best results with respect to twin pairs association. A further development would be to train a regression algorithm based on a subset of the pairs in order to see if it is possible to identify the appropriate twin based on the information from a single subject. It must additionally be underlined that the local characteristics especially when reflecting properties of neighbouring regions may be highly correlated.

6.3 WMH local information for visual scales deconstruction

As mentioned in Section 6.2.3 for the twin population, a strong correlation was observed between volumetric WMH and the Fazekas visual scale [217]. Since visual rating scales may incorporate lesion spatial distribution patterns in their depiction, the automatically derived patient-specific regional lesion burden descriptors introduced previously were then used in this context. The purpose was to improve the understanding of the spatial similarity and discrepancy between visual grading scales and the explanation for their high correlation. Here the scalar regional feature used was the local lesion frequency. After a summary of the clinical context surrounding visual scales in

Section 6.3.1, using the dataset presented in Section 5.3.6, experiments are detailed in Section 6.3.2. In light of the results reported in Section 6.3.3, an online training tool for radiologists was built in a joint effort with Ferran Prados Carrasco and is described in Section 6.3.4.

6.3.1 Clinical context

Qualitative rating scales are used throughout the domains of science with the well known Likert scale often used in psychology and social science. In the medical field, the Likert scale is for instance used to evaluate the comfort of surgeons with new surgical tools. Many other qualitative scales exist and are often used to assess for instance a subject's situation with applications going from assessment of disability level, to pain evaluation through depression status. In the imaging domain, those qualitative assessments are used for diagnostic purposes such as oncologic staging, population stratification or trial population enrichment. The neuroimaging world has seen the development of many visual scales related to the evaluation of atrophy stage [234], or to the evaluation of cerebrovascular damage with WMH [62], EPVS [235] and CMBs [236]. To be accepted and used clinically, such scales do not only have to be clinically relevant in terms of the features they assess but they must also be robust to inter- and intrarater variability.

With respect to WMH, a vast proportion of studies associating clinical findings with WMH burden have used visual rating scales. Historically, such scales were introduced as a semiquantitative way to describe the lesion burden and distribution in the brain without having to manually delineate the lesions, a task which is cumbersome and time consuming, and subject to inter- and intrarater variability. A number of visual rating scales of various complexity has been developed over time [66, 217, 218, 237–239]. Some are only applicable to magnetic resonance imaging (MRI) using either PD-weighted images or T2-weighted images, while others are designed to be used either with MRI or computed tomography (CT). The spatial information of WMH distribution incorporated in the rating scales is highly variable, ranging from whole parenchymal assessment (Manolio [240], simplified Fazekas [217]) to specific lobar lesion burden (Scheltens [218]). While spatial stratification allows for different clinical and pathophysiological explanatory pathways, the definition of the regional borders can be am-

biguous and vary from one scale to another. The question of the distinction between PVWMH and DWMH mentioned earlier is a striking example of such ambiguity. Finally, some scales have been purposefully defined for longitudinal assessment of the lesion burden whereas others are only to be applied cross-sectionally [67].

With the recent advances in the automated extraction of WMH, lesion volume has also been shown to be associated with clinical outcome, sometimes allowing for an even stronger separation between clinical subgroups [62] or relation with risk factors [241] than visual rating scales. The correlation between visual scales is considerable [65] but the heterogeneity between visual grading systems has also been put forward as a potential explanation for contradictory findings [64].

6.3.2 Data and experiments

6.3.2.1 Visual rating scales

The 82 FLAIR scans of the SABRE dataset described in Section 5.3.6 were rated by four different raters with different levels of expertise. Each rater scored the scans according to three popular visual rating scales, the Manolio scale [240], the Fazekas scale [217] and the Scheltens scale [218], that range from a global impression to more fine-grained regional scores [65]. The scales are summarised as follows:

Manolio scale [240] designed for a cardiovascular health study. The scale charaterises the lesion burden globally and ranges from 0 (absence) to 9 (highest degree) by matching to a template.

Fazekas scale [217] designed for ageing subjects in a dementia study, and the lesion rating is dichotomised between periventricular and deep WMH, assessed on a 4 point scale from 0 (absence) to 3 (highest degree), and a composite score is obtained by summing the subscales.

Scheltens scale [218] designed for ageing subjects probably affected by Alzheimer's disease. The lesion rating is defined differently according to global regions: periventricular lesions (score range 0-6), deep white matter per lobe (total score range 0-24), basal ganglia per nucleus (total score range 0-30) and infratentorial regions (score range 0-24) themselves separated in subregions. Periventricular

and deep regions are dichotomised based on the absolute distance (10 mm) to the ventricular surface.

6.3.2.2 Statistical analysis

The scores given by the different raters were averaged to produce a mean score. The average scores were correlated with the automated regional lesion burden to illustrate the spatial interactions between scores on the different scales and the frequency of lesions. In a second experiment, the individual visual scores for each one of the raters was correlated with the automated regional descriptors. With the aim of studying the degree of consistency/bias between each rater and the average, the degree of regional interactions for each rater was compared to the degree of regional interactions of the average ratings.

The global lesion burden and scale-specific aggregate regional burden estimates were used as features to predict the rating scales. A multinomial ordinal regression model was used for the regression in a stratified 2-fold cross-validation procedure with 50 repeats. Predictions were obtained with respect to the average of two, three or four raters. The ability to predict the rating scales was tested using either the global relative lesion burden or the scale-specific aggregate lesion loads.

Interrater variability was estimated as the average of the pairwise correlation between raters. Intrarater variability was estimated by repeat measurements of one single rater on a subset of 20 subjects.

6.3.3 Results

The extracted total lesion burden for the 82 subjects in this study ranged from 0.38mL to 25.28 mL (median 1.71 mL, IQR [0.81 mL 4.57 mL]). Figure 3 represents the median lesion distribution across all subjects and the corresponding IQR. It illustrates the right-left hemispherical symmetry as well as the prevalence of lesions in the perventricular zones compared to deeper layers [237], the sparing of the infratentorial regions and the tendency towards greater lesion burdens in the frontal regions [21] described in the literature.

The correlations between quantitative volumes and visual rating scales (global scores) across all raters are gathered in Table 6.2. All correlations were statistically significant with p-values < 0.0005 and not significantly different from each other. In

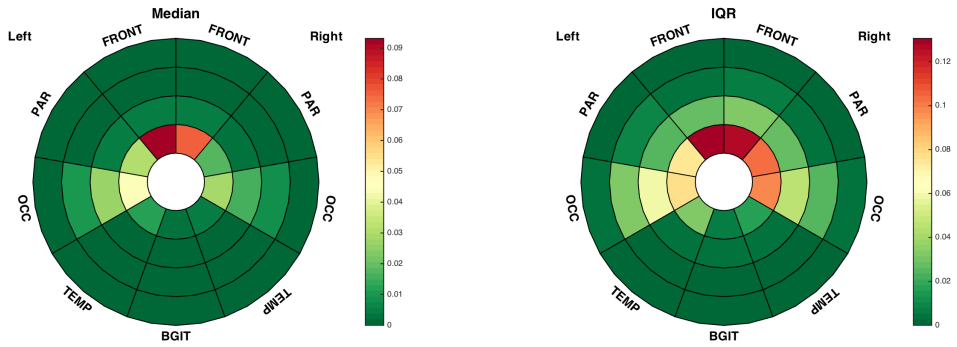


Figure 6.4: Median (left) and IQR (right) of the lesion burden frequency per local region represented in bullseyes plots.

line with the literature [58,66], there was a good agreement between the various scales. In addition, visual scales and lesion volumes were strongly related with Kendall's tau coefficients of 0.59 0.58 and 0.61 for the Scheltens, the Manolio and the Fazekas scales. The intra-rater intra-class correlation evaluated in a subset of 20 subjects were 0.73, 0.68 and 0.68 while the mean pairwise interrater ICC were 0.70, 0.80 and 0.64 for the Scheltens, Manolio and Fazekas scales respectively.

Using a similar representation as the one used in Figure 6.1, the correlations between the average Scheltens sub-scales and the regional descriptors is illustrated in Figure 6.5.

The observed correlations were found to be stronger for the subscales easier to map regionally such as the frontal and posterior periventricular regions. The clear difference in observed patterns when comparing the frontal lobe and the parietal lobe further supports the assumption that certain local features drive the visual rating process. Areas with a low density of WMH (*e.g.* temporal lobe) were not found to be highly associated with any of the regional lesion frequencies. Finally, a high degree of interaction was

	Mean	SD	Min	Max
Volume - Manolio	0.61	0.01	0.60	0.61
Volume - Fazekas	0.58	0.02	0.56	0.60
Volume - Scheltens	0.59	0.03	0.55	0.62
Manolio - Fazekas	0.72	0.02	0.71	0.75
Manolio - Scheltens	0.64	0.02	0.62	0.67
Fazekas - Scheltens	0.61	0.02	0.58	0.63

Table 6.2: Summary of Kendall's tau correlation results between global scale scores.

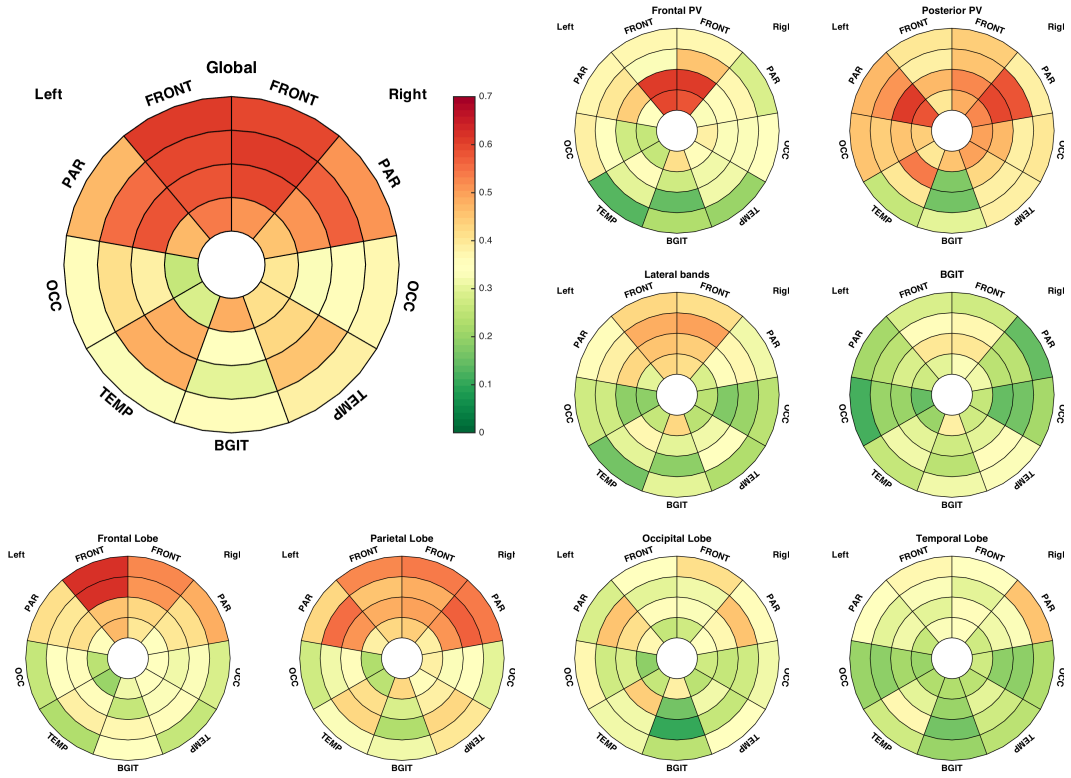


Figure 6.5: Correlation between the regional lesion loads and each Scheltens subscale. Plot titles refer to the studied regions. Note the higher correlations between the periventricular subscales and central lesion loads in the bullseyes and at the periphery of the plot for lobar scores. The bigger plot on the left represents the correlations between the global score and the local percentage of volume affected by lesions, showing that the frontal lobe had the highest overall loading.

found across all regions when correlating with the Scheltens global scale.

The correlations between each global scale and the average of four raters, as well as the difference in correlation observed between each rater and the average of the three remaining raters is presented in Figure 6.6. When focusing on the differences between a rater and the average, a pink (blue) color represents a stronger (weaker) interaction between a given rater's grades and the regional volume than the one found for the average score. Colloquially, one can interpret pink regions as over-influencing the rater's grading while blue regions are under-influencing the rater's grading when compared to the average rating. For example, in the Manolio scale grading, the influence of the three first layers of the parietal and frontal regions on rater ♯4's scores was lower than that of the average of the remaining raters for all scales, indicating that this rater should give more weight to these areas when grading.

In Figure 6.6, the three global scores (left column) show relatively similar patterns

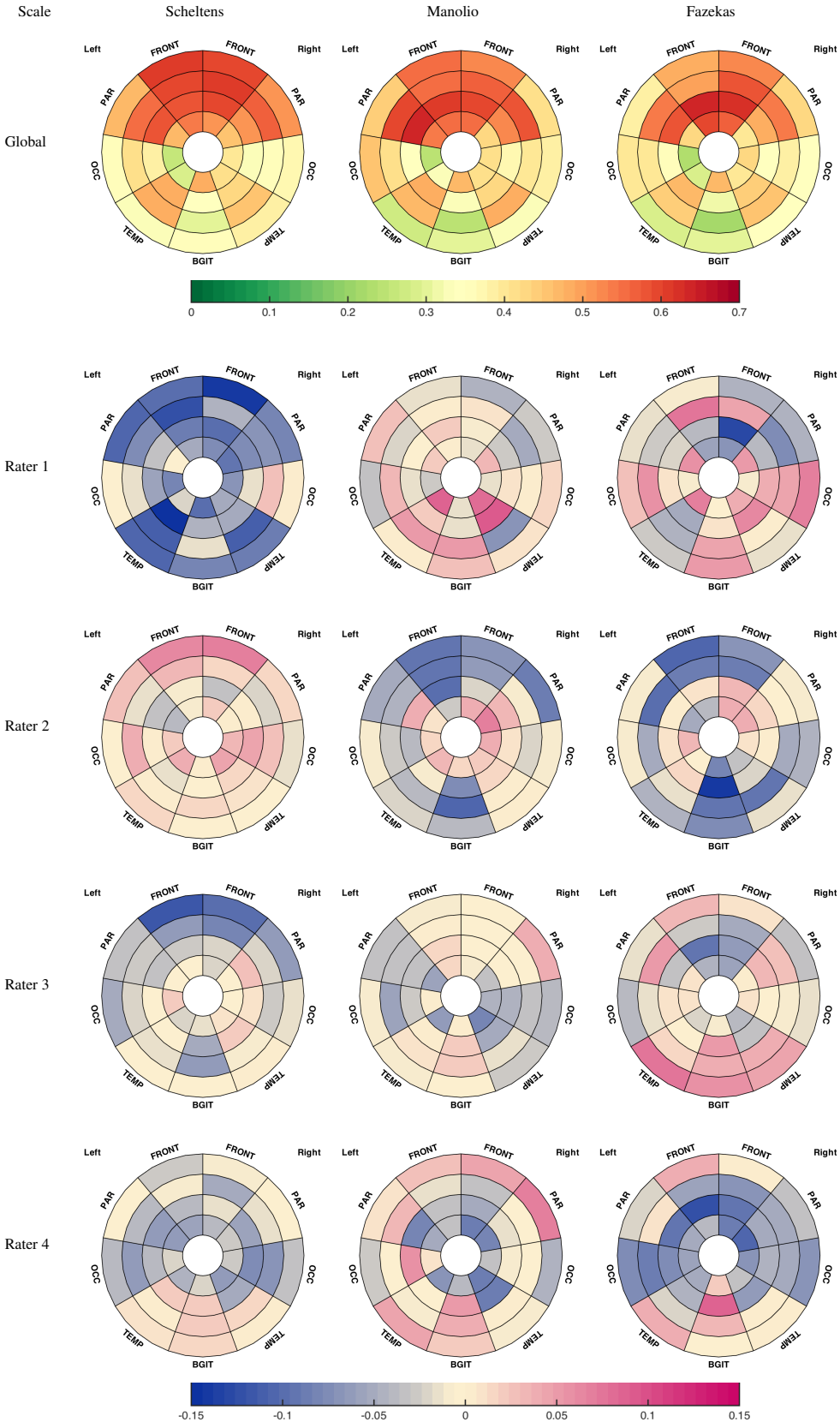


Figure 6.6: Representation for each of the studied visual scales of the Kendall's Tau correlations between the local descriptors and the global average result (1st row) and of the difference in correlation for each rater (row 2 to 5) with the correlations obtained for the average made of the three remaining raters.

		Prediction with local features			Prediction with global perc			Inter rater variability		
Scale		Pred4	Pred3	Pred2	Pred4	Pred3	Pred2	IR in 3Cons	IR in 2Cons	IR Unique
Scheltens	FPV	**0.77**	**0.75** (0.04)	**0.71** (0.07)	**0.58**	0.57 (0.02)	0.55 (0.04)	0.65 (0.15)	0.56 (0.10)	0.48 (0.14)
	LPV	**0.51**	0.47 (0.04)	0.43 (0.06)	**0.49**	0.46 (0.02)	0.42 (0.04)	0.56 (0.17)	0.45 (0.09)	0.37 (0.12)
	PPV	**0.79**	**0.75** (0.03)	**0.69** (0.06)	**0.76**	**0.72** (0.02)	**0.65** (0.05)	0.55 (0.22)	0.46 (0.13)	0.37 (0.15)
	FL	0.69	0.68 (0.02)	0.66 (0.03)	0.61	0.60 (0.02)	0.58 (0.03)	0.82 (0.07)	0.78 (0.04)	0.72 (0.05)
	PL	0.68	0.67 (0.02)	0.64 (0.03)	0.69	0.68 (0.02)	0.65 (0.02)	0.80 (0.08)	0.75 (0.04)	0.69 (0.05)
	OL	**0.64**	**0.57** (0.09)	**0.47** (0.13)	**0.59**	**0.53** (0.10)	**0.45** (0.16)	0.32 (0.29)	0.22 (0.15)	0.17 (0.16)
	TL	0.46	0.43 (0.04)	0.39 (0.07)	0.39	0.37 (0.03)	0.34 (0.06)	0.6 (0.21)	0.49 (0.13)	0.40 (0.15)
		0.81	0.80 (0.01)	0.79 (0.01)	0.80	0.80 (0.01)	0.78 (0.01)	0.88 (0.07)	0.85 (0.05)	0.80 (0.07)
	BGIT	0.53	0.51 (0.01)	0.49 (0.04)	0.50	0.49 (0.01)	0.47 (0.04)	0.81 (0.11)	0.75 (0.09)	0.69 (0.12)
	Tot	0.81	0.81 (0.00)	0.79 (0.01)	0.81	0.81 (0.00)	0.79 (0.01)	0.89 (0.06)	0.86 (0.04)	0.81 (0.06)
Manolio		**0.86**	0.85 (0.01)	0.84 (0.01)	0.84	0.83 (0.01)	0.82 (0.02)	0.88 (0.08)	0.86 (0.06)	0.82 (0.09)
	PV	**0.81**	**0.78** (0.05)	**0.73** (0.07)	**0.80**	**0.78** (0.05)	**0.73** (0.07)	0.69 (0.16)	0.63 (0.11)	0.57 (0.17)
Fazekas	DWM	**0.69**	0.68 (0.02)	0.65 (0.02)	0.68	0.67 (0.01)	0.64 (0.02)	0.75 (0.20)	0.69 (0.15)	0.61 (0.19)
	Tot	**0.82**	**0.81** (0.03)	**0.79** (0.03)	**0.80**	0.78 (0.04)	0.76 (0.04)	0.82 (0.12)	0.78 (0.09)	0.73 (0.12)

Table 6.3: Explanatory value of the automated local lesion loads.Bold font corresponds to results for which the prediction had a numerically higher or equal ICC to the training average than the mean interrater variability with the average using the same number of raters. Underlined values reflect higher correlation of the prediction with the training average than the mean pairwise ICC (last column). For the scales, the partial total refers to the sum of the Scheltens subscales excluding basal ganglia and infratentorial regions.

in the degree of regional loading with a predominant effect of periventricular zones. Compared to both the Fazekas and the Manolio scales, the Scheltens scale appears to be more homogenously distributed across all brain regions. In turn, the Manolio representation presents dominantly high correlations in periventricular regions.

The ability to explain the local and global scales based on the average gradings is gathered in Table 6.3. For all studied visual scales and subscales, the ICC between the predicted and the actual values when training on an average of two, three or four raters and using either the designed local features or the global values are calculated. Results show that: first, when predicting subscales, the use of regional burden from the same anatomical location as the sub-scale allows for better predictions than using global features. Furthermore, the ability to predict the rating scale scores appears to increase with an increase in the number of raters used to obtain the training average gradings. The correlation between average scores and prediction, based on volumetric regional predictors was higher than the interrater variability for most scales, except in regions with lower prevalence of lesions (*e.g.* temporal lobe, basal ganglia and infratentorial regions - Figure 6.5). The inter-rater correlation variance was also found to be higher than for the automated prediction model.

For all studied visual scales and subscales, the correlations between the predicted and the actual values were calculated when training on an average of two, three or four raters and using either the designed local features or the global value. The notation

Pred4 indicates for instance that the prediction was trained with the average of four raters. When appropriate (two or three raters) the results are given under the form mean (SD). The correlations are compared to the average interrater variability when correlating each rater with an average of complementary raters. Cons3 indicates for example the mean correlation between the left out rater and the average of the three other raters.

6.3.4 Creation of an online training tool in WMH visual grading scales

With the recent advance in knowledge dissemination technologies, a web-based training suite was created to help improving the precision and accuracy of raters that is now available at `http://www.cmictig.cs.ucl.ac.uk/vrt/`. The website's participant is offered a serie of 20 FLAIR scans to grade after choosing the rating scale he/she wants to be trained in as illustrated in Figure 6.7.

For each of the scans, the trainee can, with an online viewer, scroll across the images and has to decide about a grade for each of the relevant subscales. After a training session is completed, color coded local potential biases are given according to the bullseye representation along with a literary interpretation of the training thus enabling a local adjustment of the evaluation in a subsequent training. Figure 6.8 presents a view

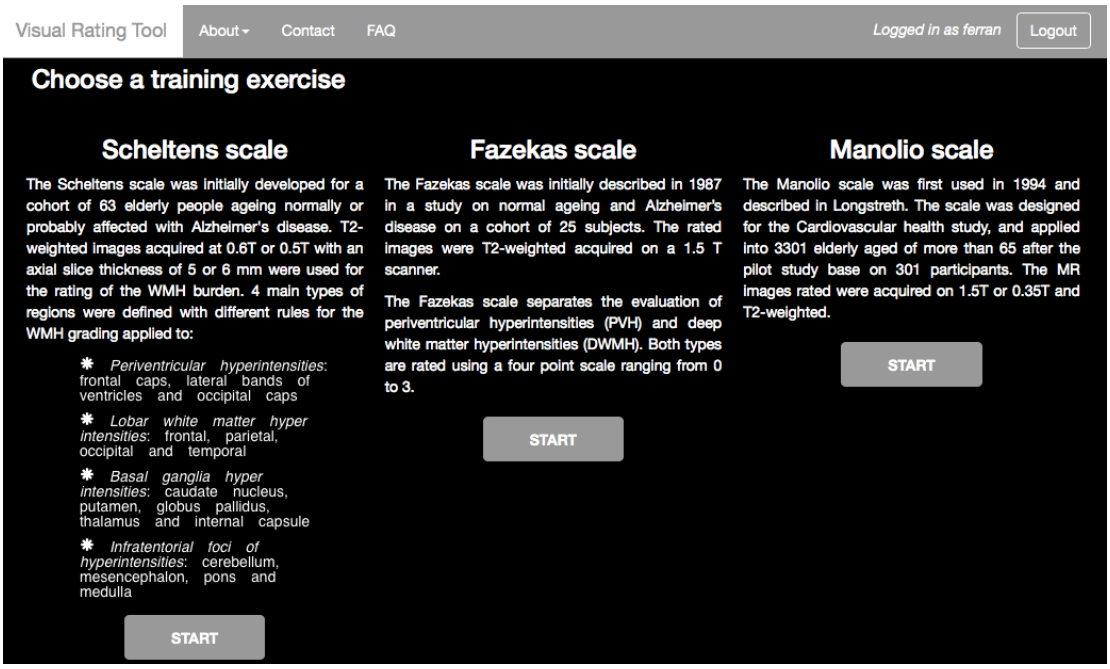


Figure 6.7: Print screen of the training system when about to choose the scale to train on.

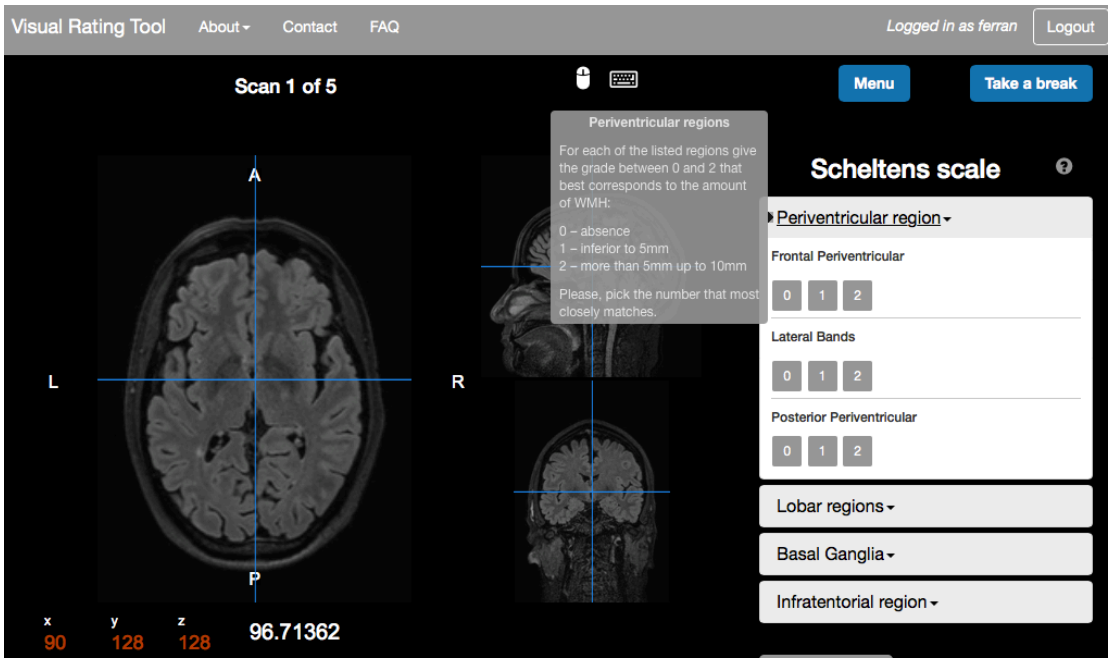


Figure 6.8: Print screen of the training system when about to rate the presented image for the periventricular subscales in the Scheltens scale. A reminder of the subscales description is always made available to the trainee.

of a page of grading for the Scheltens scale.

6.3.5 Discussion

The relevance of the regional-zonal analysis tool was here demonstrated in deconstructing visual rating scales and evaluating rater performance, for which an online training tool in visual rating is now available. Further applications may include comparison of populations, based for instance on ethnicity, vascular risk factors or clinical mode of presentation.

The regional lesion frequency features used in this work were shown to characterise both spatial similarities and differences between visual rating scales. The regional lesion loads were found to predict well the respective anatomical scores of the visual rating scales. The Manolio and the Fazekas scores showed similar spatial correlation patterns with an emphasis on the periventricular regions, while the Scheltens scores were shown to correlate in a more balanced fashion with across brain regions. The proposed data-driven approach reveals the source of discrepancies between visual rating scores, as underlined by [18,64], and can be used to inform the choice of grading scale for a clinical study or to improve the classical rating protocols.

Secondly, this new tool can illustrate the spatial source of bias between a single

rater and the consensus standard. It can show that during the rating process, some readers paid more attention to a particular region than others. The regional maps reveal the anatomical locations that bias the rating behaviour of a particular rater, which can be used to provide objective feedback. The model can therefore potentially be used as a tool for training radiologists in order to improve their rating performance and calibrate the application of visual rating scales, reducing inter and intra-rater variability. Note that the presented maps estimate the per-region rater bias under the assumption of regional independence, and does not model the correlation between regions.

Thirdly, the regional loads were shown to be predictive of the local and global average rating scales. In order to test the ability to reproduce a consensus grading, both the automated algorithm and each human rater were compared to the average of ratings. The automated prediction model performed similarly in terms of interrater variability for most regions with a reduced variance, outperforming human raters for several regions. Various factors can be put forward as limiting the model's ability to predict the average rating scores: first, an explicit choice was made regarding the regions relevant to each scale; second, the lesion burden feature used in this work does not account for the size and count criteria of the Scheltens scale, a limitation that could be mitigated by including other types of local lesion features. The proposed predictive model performed better than human raters in sub-scales with a large degree of rater disagreement, possibly due to disagreements among raters with regards to the regional definitions [18].

One of the main strengths of this study pertains to the number of raters involved in the visual grading of white matter lesions in three different scales, allowing for an exhaustive comparison between raters and scales, and for an unbiased assessment of the utility of regional features and their ability to predict the average ratings. The current study also has some limitations. The proposed method relies heavily on the accuracy of the automatic lesion segmentation and of the lobes parcellation, with segmentation errors impacting directly the analysis outcome. Also, due to ceiling and flooring effects in visual scale assessment, the assumption of normality underpinning the use of the correlation coefficient as a marker of regional influence may not hold. Finally, the relevant regions used for feature extraction were selected empirically based on the literature descriptions, possibly affecting the prediction outcome.

The quality of clinical neuroimaging has continuously improved in the past years, due to the move to higher field strength (3T) and the use of more advanced sequences. For instance, the designs of the three visual rating scales mentioned in this study were based on 2D T2 or PD weighted images obtained on 1.5T or 0.35T MR systems whereas clinical practice has evolved towards the use of FLAIR imaging and volumetric data acquisition without slice gaps. With the known increase in sensitivity, specificity and correlation with clinical outcome when using 3T images [242], changes in rating scales are expected. At higher loads, the non-linear relationship between scores and volumes highlighted by Van Straaten et al. [62], contributes to a ceiling effect of the rating scales that may explain the high inter-rater correlation observed in this work compared to the literature [66]. In those cases, using volumes rather than scales would be more appropriate, making automated classification even more relevant.

In conclusion, this application shows how local automatically extracted lesion loads can contribute to the understanding of visual rating scales, their prediction and the evaluation of raters. A web based training suite has been made available, that will expand the training potential of the local lesion assessment, helping the rater to perform local adjustments in their evaluation. Accurate semi-quantitative or quantitative assessments of WMH burden are likely to gain importance in the near future as WMH are biomarkers which can be used for assessing disease progression, therapeutic intervention (such as blood pressure lowering drugs) or risk of intervention (carotid stenting) and the bullseye plots will help to visualise associations with risk factors or differences between populations.

6.4 Discussion

This chapter focused on the derivation and application in various contexts of a systematic patient-specific regional representation of the lesion spatial distribution. In order to account for atrophy and global ageing changes, the otherwise dominant absolute distance to the ventricles was abandoned in favour of a relative distance measure derived from the application of Laplace equation. Regions were delineated based on the propagation of lobar parcellation and discretisation of this distance map. They are therefore dependent on the quality of the segmentation and on the choices regarding the discretisation. At the population level, this systematic spatial information on the lesion

distribution appeared to be beneficial to correctly associate twin pairs as demonstrated in Section 6.2.3. Furthermore, its relevance was demonstrated in the context of the description and prediction of visual rating scales (Section 6.3.3). In fact, it allows the representation of tridimensional information in a bidimensional systematic way. In the presented cases, volumetric characterisations were mostly used but other traits, such as intensity, cardinality or shape may also prove pertinent to be studied locally in the case of WMH. The systematic location scheme for lesion spatial distribution was used in the two presented applications in a perspective of synthesis of information. Other applications could include the spatial characterisation of individual lesions or the longitudinal evolution of the spatial distribution. The latter requires however a robust longitudinal segmentations of WMH, task that will be examined in the following chapter.

Chapter 7

Extension to longitudinal studies

7.1 Contextualisation

From afar or close, the concept of longitudinal change is inherent to any investigation related to ageing or evolving pathologies and the problem of assessing changes is at the core of such studies. In a research-focused perspective, assessment of changes are key points in further understanding the pathways of a biological process. Furthermore, when implementing clinical trials, measures of change are often used as surrogate end-points to assess the effect of a tested drug. Thus, robust and accurate longitudinal measures of imaging biomarkers are necessary. In the case of WMH, even though most studies are cross-sectional, progression is observed with time and has been largely correlated with detrimental clinical outcome [243] in processing speed [244], executive and motor function [245]. Risk factors including variability in systolic blood pressure [246] and current smoking [41] have also been shown to relate to the progression of WMH. Moreover, the baseline lesion load of WMH appears repeatedly as a strong predictor of WMH progression [3, 247]. Dedicated visual rating scales have been developed for the assessment of lesion changes but volumetric analysis applied cross-sectionally has been observed to be more reliable than visual rating scales [248]. Although semi automated volumetric assessments may tend to overestimate the change when the scans are observed side by side [249], differences across automated methods in detected volume change may also be quite high [250].

In the ageing process, automatic measurements of change such as atrophy of the whole brain, specific structures or tissues have been greatly discussed and guidelines have been defined to warn against and avoid generating bias [251]. As underlined by

Reuter et al. [252], bias in longitudinal measurements can be introduced in different ways ranging from the acquisition, due to changes in protocols and scanners, to processing when a specific time point in the serial images is handled differently compared to the others. In most longitudinal frameworks the main principle is to make use of the inherent within subject correlation of measurements in order to reduce the variability due to noise. Not accounting for this correlation may prove inefficient and lead to increased sample sizes as underlined by Elliott et al. [253]; including all time points is stated as a point for an improvement in quantification by Vrenken et al. [102].

Specific volumetric longitudinal strategies devoted to lesion in the WM have for now been developed in the case of MS. Existing strategies use for instance intensity subtraction between pair of images [254, 255] but can be biased by errors in the registration. Similarly, studying the deformation field may not be enough to account for appearing lesions [253] when registration errors occur. Methods relying on the analysis of the difference between registered time points images, as Rey et al. do [256], may be further hindered by other volumetric changes occurring between the time points such as atrophy or oedema.

Furthermore, in studies with long-term follow-up, in which the drop-off rate can be high, (*e.g.* in age-related studies), being able to handle different numbers of time points without constraint is necessary. In the context of age-related WMH, the progressive characteristic of the damage can be taken as an argument for looking progressively at image pairs as performed by Bosc et al. [257]. However, noise and artefacts, prevalent in ageing or demented population, may affect methods based on progressive inference [258]. Conversely, other solutions based on image averaging and model building may prove advantageous.

7.2 Methods

To account for all these reported challenges, the use of average images to guide the processing of longitudinal data has been promoted by Reuter et al. [259]. Thus, the solution developed here consists first in creating average data summarising the per-subject information contained at all time points (Section 7.2.1), followed by the derivation of an appropriate Gaussian mixture model (GMM) (Section 7.2.2) that will finally be used to constrain the data model at each time point (Section 7.2.3). At that stage, lesion seg-

mentation can be obtained in a post-processing step, similarly to what has been detailed in Chapter 5. The main assumption of this work, presented at the BAMBI workshop at MICCAI 2015 is that despite changes related to atrophy and lesion expansion, each subject's brain will retain most of its morphological characteristics through time.

In addition to the notations introduced in Chapter 4, in the following, the subscript τ denotes the time point and GW the groupwise average. With N the number of voxels and D the number of modalities, images intensities are vectorised into $Y^{(d)} = \{y_{d1}, \cdots, y_{dn}, \cdots y_{dN}\}$ with y_{dn} the intensity at voxel n of modality d, so that

$$\mathbf{Y} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(D)} \end{pmatrix}.$$

7.2.1 Creation of an intra-subject average image

In order to build the average appearance model, two main components linking the individual images to the average space are needed: a spatial transformation and an intensity transformation. An intensity matching between images is needed to account for changes in contrast, MR scanning variations and artefacts. These transformations are obtained through an iterative process, proved to limit bias towards a specific time point.

In order to avoid unrealistic spatial deformations, affine transformations roughly aligning the images are first applied before considering non-rigid transformations to obtain the final spatial transformations $T_{\tau \to \mathrm{GW}}$. At each iteration, the intensities of the images spatially transformed to the GW space are mapped to the intensities of the current average image using a polynomial fit of degree 2 for each modality. In order to avoid the mapping to be biased by the presence of lesions or artefacts, it is possible to exclude the most obvious outliers to obtain the mapping coefficients. More formally, the intensity matching and the resulting mapping coefficients $\hat{h}_\tau$ for one modality on the bias field corrected log-transformed intensities $\hat{Y}_\tau$ can be expressed as

$$\hat{h}_\tau = \underset{h_\tau}{\operatorname{argmin}} \parallel A\left(\hat{Y}_\tau(T_{\tau \to \mathrm{GW}}))\right) h_\tau - \hat{Y}_{\mathrm{GW}} \parallel^2$$

where $A\left(\hat{Y}_\tau(T_{\tau\to\mathrm{GW}})\right)$ is the polynomial matrix transformation of $\hat{Y}_\tau(T_{\tau\to\mathrm{GW}})$ such that

$$A(\hat{Y}) = \begin{pmatrix} 1 & \hat{y}_1 & \hat{y}_1^2 \\ \vdots & \vdots & \vdots \\ 1 & \hat{y}_N & \hat{y}_N^2 \end{pmatrix}.$$

The steps to create an average appearance model are:

Step 1 Register each of the individual time points to the current average image.

Step 2 Map the relevant intensities of each resampled image to the current image using a polynomial fit of degree 2.

Step 3 Average all resampled and intensity transformed images to create the new current average image.

Step 4 Go back to step 1.

With this set up the loop is performed five times: the first time estimating a rigid transformation then two affine transformations before allowing for a non-rigid registration at the last two iterations. It has proved empirically to be enough to reach a reasonable convergence.

7.2.2 Model selection

After creating the average appearance model, patient-specific tissue priors and brain mask are obtained using the GIF pipeline [198]. In this method, label-fusion is used to propagate the tissue segmentation, and the anatomical tissue priors ($\mathbf{A}$) are obtained as the output of the fusion step. BaMoS (cf Chapter 4) is applied on the average image with a marginal modification to the initialisation stage. After the initial EM convergence that uses flat inlier/outlier priors, uniform distributions for the outlier tissue classes and a unique Gaussian component for each inlier class, a typicality map, that represents the ability of the inlier part of the model to represent each voxel, is derived according to the method detailed by Van Leemput et al. [79]. The typicality value for a

given voxel n follows the expression

$$t_n = \sum_{j=1}^{J} p_{nj} \frac{\mathscr{G}\left(\mathbf{y}_n \middle| \theta_{I_j}\right)}{\mathscr{G}\left(\mathbf{y}_n \middle| \theta_{I_j}\right) + \dfrac{1}{\sqrt{(2\pi)^D \left|\Lambda_{I_j}\right|} \exp\left(-\dfrac{1}{2}\kappa^2\right)}},$$

where $p_{nj} = p_{nI_j} + p_{nO_j}$ and $\kappa = 3$.

In this expression, the level of outlierness accounts for the variability in the shape of covariance matrices, and therefore balances the sensitivity of the outlier detection between tissue classes. This typicality map, smoothed with Gaussian filtering is used instead of the flat priors to perform BaMoS model selection framework. Once the final GMM is obtained, the atlases are replaced by the smoothed result of the tissue segmentation. The full model (number of Gaussian components, parameters and atlases) is used to constrain the segmentation at each time point.

7.2.3 Constraint over time point

Three main components are needed to perform the individual time point segmentation based on the model derived for the average image. In addition to the model on the average, both the spatial and the intensity transformations relating each time point to the average space are required. The statistical patient-specific atlases are transformed into the individual time points while the bias-field corrected intensities of the time point image are transformed towards the intensity space of the average image by the application of the polynomial transformation $M_{\hat{\mathbf{h}}_\tau}$ described by the coefficients $\hat{\mathbf{h}}_\tau$ on the data $\hat{\mathbf{Y}}_\tau$. The introduction of constraints over the time point model parameters, through priors based on the average model, is made possible by the intensity transformation. With these constraints, the time point model is simply obtained by a parameter optimisation of the model on the average through an EM procedure without allowing the complexity structure (number of components per tissue class) to evolve. The diagram representing the constraints over the time point is displayed in Figure 7.1. The introduction of priors over the parameters to constrain the model is obtained using a normal distribution for the mean and an Inverse-Wishart distribution for the covariance in order to maintain a possible decoupling for the parameters optimisation. As such the expectation step is

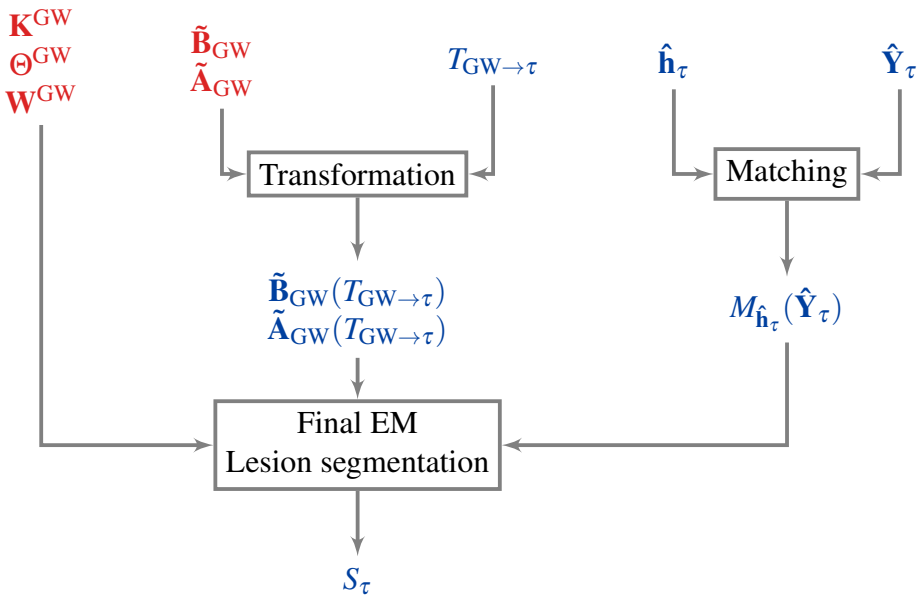


Figure 7.1: Diagram of the constraints operated on each time point based on the model on the average. Groupwise elements are in red while time point specific entities are in blue.

not modified but the M-step consists in maximising the following:

$$\mathscr{Q}\left(\boldsymbol{\Xi}^{\tau}_{K^{\mathrm{GW}}}\middle|\boldsymbol{\Xi}^{\tau^{(t-1)}}_{K^{\mathrm{GW}}}\right)=\mathbb{E}_{\mathbf{Z}_\tau|\hat{\mathbf{Y}}_\tau,\boldsymbol{\Xi}^{\tau^{(t-1)}}_{K^{\mathrm{GW}}},\hat{\mathbf{h}}_\tau}\log\left[f\left(\hat{\mathbf{Y}}_\tau,\mathbf{Z}_\tau\middle|\boldsymbol{\Xi}^{\tau}_{K^{\mathrm{GW}}},\hat{\mathbf{h}}_\tau\right)\cdot f\left(\boldsymbol{\Xi}^{\tau}_{K^{\mathrm{GW}}}\middle|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right)\right].$$

The distribution $f\left(\boldsymbol{\Xi}_{K^{\mathrm{GW}}}\middle|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right)$ in which the script τ has been dropped for notation convenience is expressed as

$$f\left(\boldsymbol{\Xi}_{K^{\mathrm{GW}}}\middle|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right)=\prod_{l\in I,O}\prod_{j=1}^{J}\prod_{k=1}^{K_{l_j}}\mathscr{G}\left(\boldsymbol{\mu}_{j_{l_k}}\middle|\boldsymbol{\mu}^{\mathrm{GW}}_{l_{j_k}},\Lambda^{\mathrm{GW}}_{l_{j_k}}\right)\mathscr{W}^{-1}\left(\Lambda_{l_{j_k}}\middle|\tilde{N}\Lambda^{\mathrm{GW}}_{l_{j_k}},N\right),$$

where $\mathscr{G}$ refers to a normal distribution and $\mathscr{W}^{-1}$ to an Inverse-Wishart distribution with $\tilde{N}=N+D+1$. The differentiation with respect to $\boldsymbol{\mu}_{l_{j_k}}$ then results in

$$\begin{aligned}\frac{\partial\mathscr{Q}}{\partial\boldsymbol{\mu}_{l_{j_k}}}&=\sum_{n=1}^{N}p_{nl_{j_k}}\frac{\partial}{\partial\boldsymbol{\mu}_{l_{j_k}}}\log\mathscr{G}\left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n)\middle|\boldsymbol{\mu}_{l_{j_k}},\Lambda_{l_{j_k}}\right)+\frac{\partial}{\partial\boldsymbol{\mu}_{l_{j_k}}}\log\mathscr{G}\left(\boldsymbol{\mu}_{l_{j_k}}\middle|\boldsymbol{\mu}^{GW}_{l_{j_k}},\Lambda^{GW}_{l_{j_k}}\right)\\&=\sum_{n=1}^{N}p_{nl_{j_k}}\left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n)-\boldsymbol{\mu}_{l_{j_k}}\right)\Lambda^{-1}_{l_{j_k}}-\left(\boldsymbol{\mu}_{l_{j_k}}-\boldsymbol{\mu}^{GW}_{l_{j_k}}\right)\Lambda^{GW\,-1}_{l_{j_k}}.\end{aligned}$$

Solving for $\boldsymbol{\mu}_{l_{j_k}}$ that annuls it then leads to

$$\boldsymbol{\mu}_{l_{j_k}} = \left(\sum_{n=1}^{N} p_{nl_{j_k}} M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) \Lambda_{l_{j_k}}^{-1} + \boldsymbol{\mu}_{l_{j_k}}^{GW} \Lambda_{l_{j_k}}^{GW\,-1} \right) \cdot \left(\sum_{n=1}^{N} p_{nl_{j_k}} \Lambda_{l_{j_k}}^{-1} + \Lambda_{l_{j_k}}^{GW\,-1} \right)^{-1}.$$

Similarly the derivation with respect to $\Lambda_{l_{j_k}}$ can be expressed as

$$\begin{aligned}
\frac{\partial \mathscr{Q}}{\partial \Lambda_{l_{j_k}}} =& \sum_{n=1}^{N} p_{nl_{j_k}} \frac{\partial}{\partial \Lambda_{l_{j_k}}} \log \mathscr{G} \left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) \middle| \boldsymbol{\mu}_{l_{j_k}}, \Lambda_{l_{j_k}} \right) \\
&+ \frac{\partial}{\partial \Lambda_{l_{j_k}}} \log \mathscr{W}^{-1} \left(\Lambda_{l_{j_k}} \middle| \tilde{N} \Lambda_{l_{j_k}}^{GW}, N \right) \\
=& \frac{1}{2} \left[\sum_{n=1}^{N} p_{nl_{j_k}} \left(-\Lambda_{l_{j_k}}^{-1} + \Lambda_{l_{j_k}}^{-1} \left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) - \boldsymbol{\mu}_{l_{j_k}} \right)^T \left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) - \boldsymbol{\mu}_{l_{j_k}} \right) \Lambda_{l_{j_k}}^{-1} \right) \right. \\
&\left. + \left(-\tilde{N} \Lambda_{l_{j_k}}^{-1} + \Lambda_{l_{j_k}}^{-1} \tilde{N} \Lambda_{l_{j_k}}^{GW} \Lambda_{l_{j_k}}^{-1} \right) \right].
\end{aligned}$$

Finding the solution for $\Lambda_{l_{j_k}}$ that nulls the differentiation obtained above consists of resolving

$$\begin{aligned}
\left(\tilde{N} + \sum_{n=1}^{N} p_{nl_{j_k}} \right) \Lambda_{l_{j_k}}^{-1} &= \left(\sum_{n=1}^{N} p_{nl_{j_k}} \Lambda_{l_{j_k}}^{-1} \left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) - \boldsymbol{\mu}_{l_{j_k}} \right)^T \left(M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) - \boldsymbol{\mu}_{l_{j_k}} \right) \Lambda_{l_{j_k}}^{-1} \right) \\
&\quad + \tilde{N} \Lambda_{l_{j_k}}^{-1} \Lambda_{l_{j_k}}^{GW} \Lambda_{l_{j_k}}^{-1} \\
\tilde{N} + \sum_{n=1}^{N} p_{nl_{j_k}} &= \Lambda_{l_{j_k}}^{-1} \left[\sum_{n=1}^{N} p_{nl_{j_k}} \Delta_{l_{j_k}} + \tilde{N} \Lambda_{l_{j_k}}^{GW} \right] \\
\Lambda_{l_{j_k}} &= \frac{\sum_{n=1}^{N} p_{nl_{j_k}} \Delta_{l_{j_k}} + \tilde{N} \Lambda_{l_{j_k}}^{GW}}{\tilde{N} + \sum_{n=1}^{N} p_{nl_{j_k}}}.
\end{aligned}$$

With $\Delta_{l_{j_k}}^{(t)}$ the weighted covariance matrix, the parameters updated during the maximisation step are modified into

$$\begin{aligned}
\boldsymbol{\mu}_{l_{j_k}}^{(t)} = & \left(\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n) \Lambda_{l_{j_k}}^{-1(t-1)} + \boldsymbol{\mu}_{l_{j_k}}^{\mathrm{GW}} \Lambda_{l_{j_k}}^{\mathrm{GW}\,-1} \right) \\
& \cdot \left(\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} \Lambda_{l_{j_k}}^{(t-1)^{-1}} + \Lambda_{l_{j_k}}^{\mathrm{GW}\,-1} \right)^{-1}
\end{aligned}$$

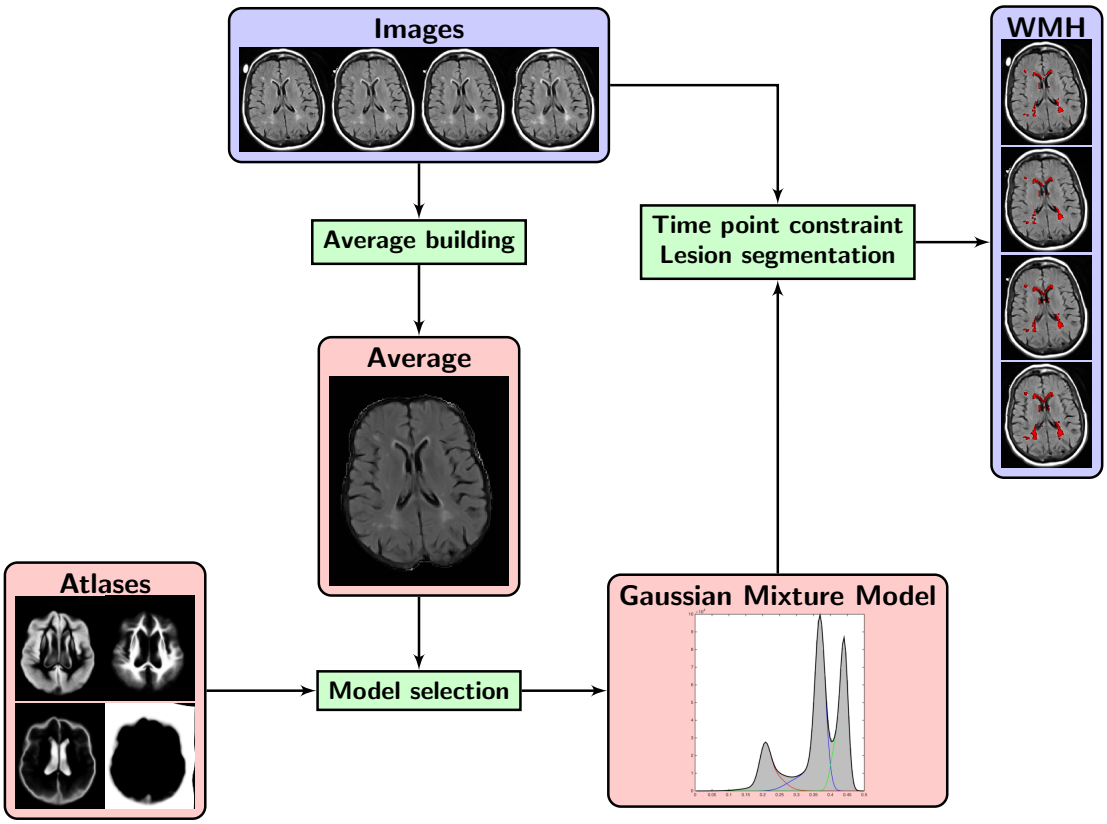


Figure 7.2: Scheme of the main aspects of the longitudinal framework.

$$\Lambda_{l_{j_k}}^{(t)} = \frac{\tilde{N}\Lambda_{l_{j_k}}^{\mathrm{GW}} + \sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} \Delta_{l_{j_k}}^{(t)}}{\sum_{n=1}^{N} p_{nl_{j_k}}^{(t)} + \tilde{N}}.$$

The global scheme with the three main steps of the longitudinal framework (average creation, model selection, time point constraint) is displayed in Figure 7.2.

7.3 Validation through simulation

7.3.1 Lesion simulator

7.3.1.1 Context

Due to the difficulty to obtain reliable, robust and reproducible ground truth from manual segmentations, a potential alternative is the use of simulated data. In the majority of the cases, an MRI simulator uses the pulse sequence parameters with the inclusion of additional artefacts such as bias field and noise and produce completely synthetic images based on tissue maps characteristics. Typically, the Brainweb project is such an

example and is recommended as an initial step in the validation of lesion segmentation algorithms [50]. As underlined by Lladó et al. [50] and Prastawa et al. [260], these simulators do not allow for much variability and realism in the depiction of pathology. In the case of Brainweb, only three lesion models of MS type lesions are available. Using statistical texture, Prastawa et al. [260] have simulated contrast-enhanced MR images of brain tumours. Closest to a longitudinal simulation, Melhem et al. [261] have developed a simulator of healthy tissues and a lesion of varying size whose intensity characteristics were derived from a single clinical scan. The simulation was designed to evaluate the limit at which change in a lesion becomes noticeable for trained observers. However, global longitudinal simulations are yet to be developed. In the lesion simulator developed here, clinical images are used as a basis both for healthy tissue and typical lesion location. Similarly to the synthetic images created for validation purposes in the studies of Jack et al. [103] and Gibson et al. [156], lesions are transferred on the background clinical images. Since the transfer is based on probabilistic maps, the applied lesion maps can then be adapted and evolution patterns simulated.

7.3.1.2 Cross-sectional simulation of images with lesions

In line with the synthetic image simulator detailed by Jack et al. [103], two sets of data are used to simulate lesions: a receiving set, composed of subjects with minimal to no WMH, and a donating set, composed of images with noticeable WMH lesion load and their associated probabilistic lesion segmentation L. The process of simulating lesions involves the spatial transformation of the lesions from the donating set to the receiving set and the creation of appropriate lesion intensities.

After registration, smoothing and normalisation of the donating lesion map to the receiving set, the probability maps are transformed into intensity maps by drawing samples of Gaussian distributions with parameters derived from the data distribution of the receiving set and using the probabilistic map to weight them.

To account for variation over time of scanner characteristics and subject positioning, random bias field and rigid transformations are applied to the images. The bias field, modelled as a linear combination of polynomial basis functions (cf Section 3.2.1) is obtained by randomly choosing the basis coefficients.

Given a probabilistic lesion map L, the Gaussianly sampled lesion intensities G,

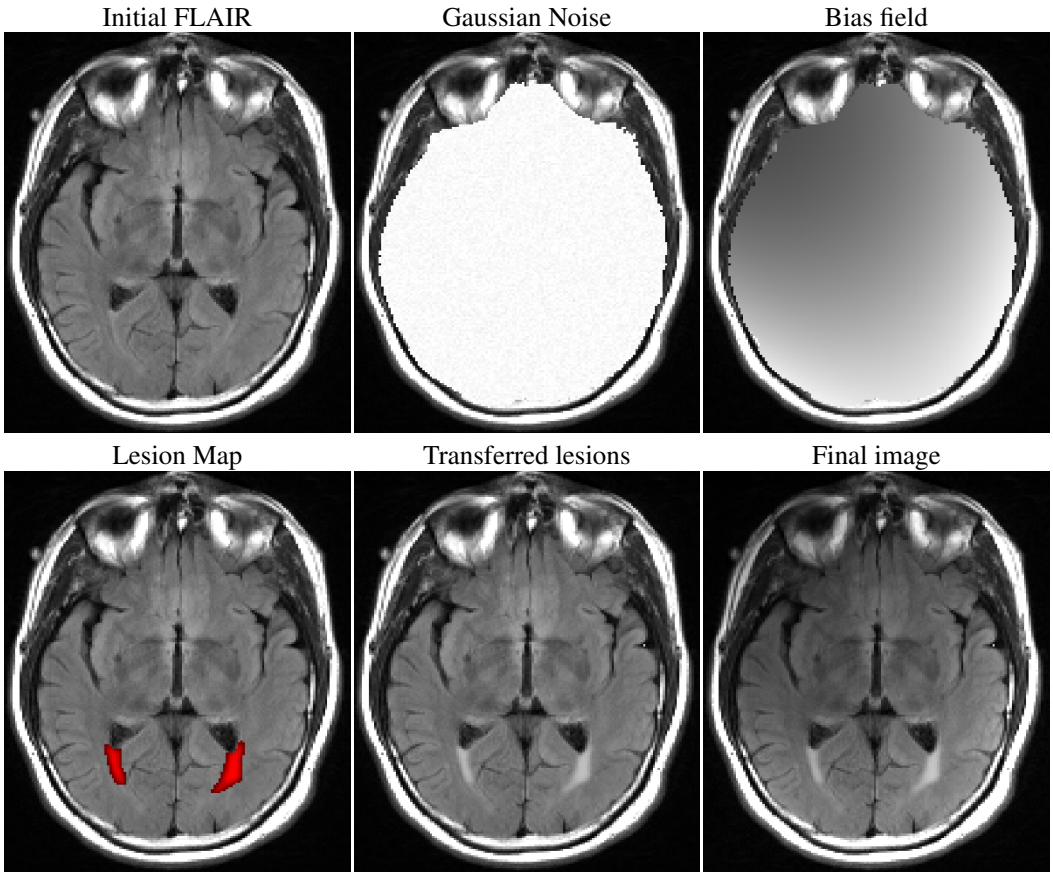


Figure 7.3: Different steps in the production of the simulated image with lesion from the initial lesion free image to the final image. In this case, no rigid transformation is applied.

the log-transformed bias field BF, the rigid transformation R and the initial image I, the final simulated image S can be expressed at voxel n by

$$S_n = R\left(\exp(BF_n)\left(L_n G_n + (1 - L_n) I_n\right)\right).$$

An example of the sequential intermediate steps of the production of the image with inpainted lesion starting from an initial image with minimal lesion load is presented in Figure 7.3.

7.3.1.3 Longitudinal simulation

In order to ensure the realism of the lesion evolution and avoid extension of the lesions onto unwanted tissues, the simulation process starts at the point of maximum lesion load. Furthermore, if the ventricle segmentation is known, the periventricular nature of lesions can be maintained.

To simulate smaller lesions, the initial propagated WMH load is modified by

thresholding the probabilistic lesion segmentation at a certain value X, followed by a normalisation step, *i.e.* $L_n^{\text{new}} = (L_n - X)/X \ \forall L_n > X$, with L being the original lesion probability map. The value of X is chosen to produce an exact volumetric reduction in WMH of D. As L^{new} can contain non biologically plausible hard edges, it is then Gaussianly smoothed. Due to the non-volume-preserving nature of the Gaussian smoothing process, L^{new} is finally re-mapped to have an exact reduction in WMH volume of D trough a piecewise linear transformation.

Defining p_b such that

$$\sharp L_S | L_S > p_b = \sharp L | L > 0.5 - D,$$

the following system to define the two linear mapping is solved:

$$\left.\begin{array}{ll} b_1 & = 0 \\ a_1 \cdot p_b + b_1 & = 0.5 \end{array}\right\} \text{if } L_S < p_b \qquad \left.\begin{array}{ll} a_2 \cdot p_b + b_2 & = 0.5 \\ a_2 + b_2 & = 1 \end{array}\right\} \text{if } L_S > p_b$$

In case the volume by which to decrease the lesion volume is higher than the remaining lesion volume itself, no change is applied to the lesion map. This is done to ensure a minimal amount of WMH as usually observed in ageing population. Considering an evolution pattern with more than two time points, the changes occur iteratively, *i.e.* a change in volume is applied on the last built lesion map. Although this would limit the evolution to monotonic decreasing patterns, the lesion maps order can be reshuffled.

The lesion simulator flowchart is displayed in Figure 7.4. An example of the outcome of the lesion simulator is presented in Figure 7.5 on which the same slices of the FLAIR image (before rigid transformation and bias field application) is presented at four time points of the lesion evolution with a non linear change of 15% of the volume at each time point.

As seen in Chapter 5, the use of a manual gold standard for the validation of automated algorithm is questionable and obtaining valuable manual segmentations is a long and difficult process. Additional measures of segmentation robustness and validity as those presented in Section 5.2.4 can be of interest but do not reflect the relevance of

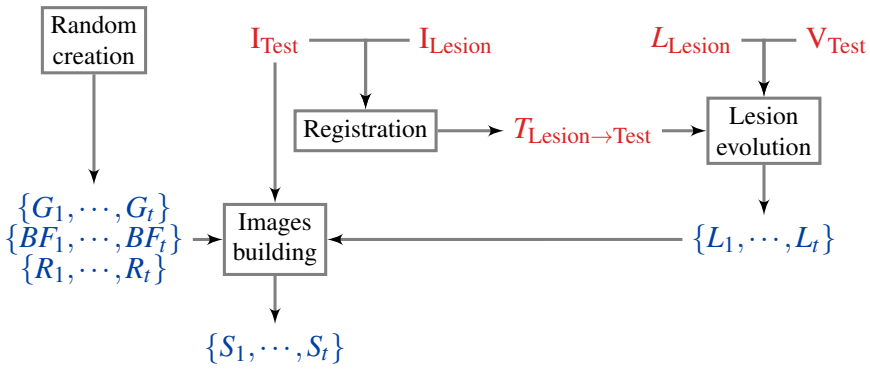


Figure 7.4: Representation of the lesion simulator flow chart. I refers to the initial images, L to probabilistic lesion maps, V to the ventricle segmentation, G to the Gaussian intensity sampling, BF to the randomly generated bias field, R to the rigid transformations applied during the simulation and S to the simulated images. The red elements correspond to unique images whereas blue entities correspond to sets of multiple elements.

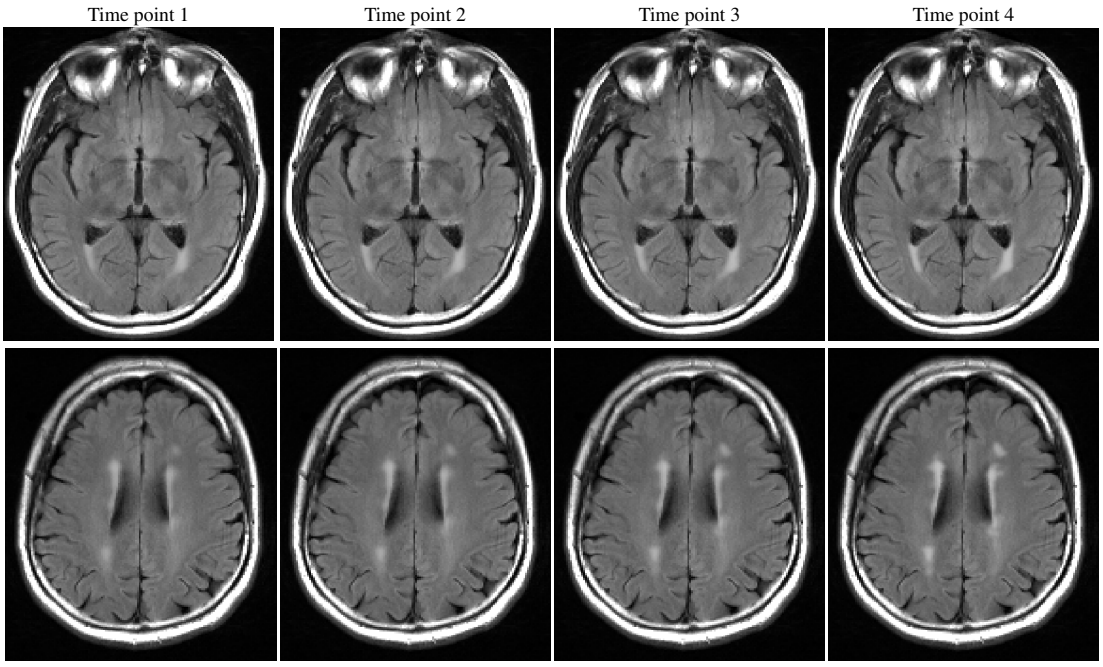


Figure 7.5: Results of the lesion simulator before application of the bias field and of the affine transformation for four time points with a volume change of 15% per step on two slices. For realism purposes and to simulate an overall increase in the lesion loads, the order in which the images are simulated is reversed.

the segmentation. In the case of longitudinal assessment, given the variability inherent to manual segmentations, classical overlap assessments are not practical. The simulator described above was used as an alternative to those issues so as to test the longitudinal framework in different conditions of longitudinal evolution and compared to the cross-sectional versions of BaMoS.

7.3.2 Experiments

7.3.2.1 Evaluation schemes

Since this simulator was designed to account for longitudinal changes, different evolution patterns with varied maximum lesion loads and diverse number of time points were tested.

The database used for WMH simulation comprised of 5 donating images and 17 receiving images. 4 patterns were simulated using the following:

Linear_500 : Linear reduction of 500 mm^3 per step, spanning 6 time points.

Linear_750 : Linear reduction of 750 mm^3 per step, spanning 4 time points.

NonLinear_5 : Non-linear reduction of 5% per step, spanning 6 time points.

NonLinear_15 : Non-linear reduction of 15% per step, spanning 4 time points.

Although the progressions are simulated by reducing the lesion load, for clinical realism and illustration purposes, the time points were then reordered to simulate a progressive increase in lesion load.

For each of these progression schemes, two additional plateauing patterns were added to test for longitudinal bias:

Flat_High 1 time point with highest load was added to form a high plateau.

Flat_Low 2 time points with lowest load were added to form a low plateau.

Finally, to simulate treatment effect, composite patterns were created using the linear patterns in order to simulate changes in the slope:

Treatment One increase step of 750 mm^3 followed by two steps with an increase of 500 mm^3 each.

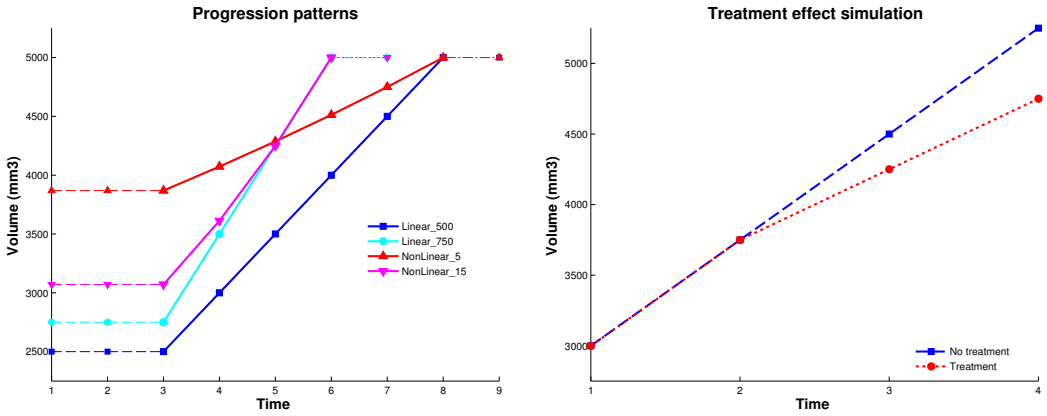


Figure 7.6: Left). Example of the four tested evolution patterns. The dashed horizontal lines represent the plateauing experiments at either high (Flat_High) or low (Flat_Low) load. Right) Example of the combination of two linear patterns to model a treatment related change.

No treatment 3 steps with an increase of 750 mm^3 per step.

Figure 7.6 left plots the four typical evolution paths with different minimum loads and their associated plateauing versions while Figure 7.6 right presents the combination of linear patterns to simulate the treatment effect case.

7.3.3 Segmentation assessment

The longitudinal framework (Long) was compared to the cross-sectional application of BaMoS both in its original version (Cross) [262] and its sensitivity enhanced variant (Cross+), *i.e.* when using the typicality map to estimate the outlier atlas. As the receiving images used in the simulation can contain trace amounts of lesions, the region where all methods agreed to the presence of lesions for the time point with minimal lesion load was excluded from the analysis, both in terms of volume and overlap.

Ground truth (GT) lesion segmentations are thus the simulated lesion probability maps corrected for the baseline lesion segmentation intersection of all given methods. Statistics of the ground truth volumes are presented in Table 7.1. Those corrected dif-

	Linear_500	Linear_750	NonLinear_5	NonLinear_15
Mean	2645	2793	2871	2510
SD	2594	2619	2519	2307
Median	2379	2542	2881	2206
IQR	[188 3941]	[314 4176]	[467 4040]	[401 3745]

Table 7.1: Summary of the ground truth volumes (Lesion probability map - intersection of baseline segmentations) across the different evolution patterns.

	DSC	TPR	AvDist	OE/TotF	OEFP/FP	OEFN/FN	FP/TotF
Linear_500	64.9 [28.1 77.1]	83.3 [65.9 90.6]	2.07 [1.00 9.87]	80.5 [48.3 90.4]	72.1 [40.7 86.8]	96.6 [88.1 100]	82.7 [59.8 91.7]
Linear_750	66.3 [35.2 77.2]	80.2 [56.9 88.5]	2.00 [0.86 6.06]	80.2 [57.4 90.6]	70.3 [45.7 87.0]	94.0 [81.0 98.0]	78.9 [56.4 87.9]
NonLinear_5	66.1 [40.5 76.0]	81.0 [64.7 88.4]	1.90 [0.96 5.51]	82.4 [67.3 90.4]	74.1 [54.4 87.0]	97.4 [91.9 99.6]	77.1 [55.7 86.9]
NonLinear_15	64.4 [41.4 76.5]	78.6 [64.5 86.3]	2.02 [0.98 6.28]	79.1 [60.4 89.9]	70.4 [51.0 85.1]	95.2 [87.8 98.6]	75.2 [57.7 85.3]

Table 7.2: Segmentation assessment table for the longitudinal framework according to the different strategies of evolution. Definitions of the assessment measures are given in Section 5.2. AvDist is given in mm and all the other measures in %.

ferences were finally compared in terms of Dice score coefficient (DSC), true positive rate (TPR) and average distance (AvDist) as defined by Styner et al. [159] and their statistics calculated over subjects and time points. The origin of the errors was further investigated differentiating false positive (FP) and false negatives (FN), outline (OE) and detection error (DE) as introduced by Wack et al. [202] and detailed in Section 5.2. Due to the non normality of the differences, when comparing across methods, paired Wilcoxon tests were used while two sample tests for unmatched data were used when comparing across evolution patterns.

7.3.4 Results

7.3.4.1 Evolution patterns and bias

The assessment across the evolution patterns are presented in Table 7.2. There was no significant statistical difference relative to the differences in DSC for the pairwise Wilcoxon tests performed but a trend towards better scores for patterns with lower ranges of change was observed (NonLinear_5 and NonLinear_15). Confirming the observations of relationships between lesion load and DSC mentioned in Section 5.2.2, lower loads (Linear_500) showed a trend towards lower DSC.

Possible bias introduced by a flat WMH load at the lowest (Flat_Low) or at the highest (Flat_High) end of the progression period was evaluated on the common time points between the sets. The results for this experiment are presented in Table 7.3 emphasising the stability of the method when including plateauing time points.

7.3.4.2 Comparison between cross-sectional and longitudinal methods

In order to compare the proposed longitudinal version of BaMoS with the cross-sectional methods, the assessment measures were calculated for the 1360 baseline cor-

	DSC	TPR	AvDist	OE/TotF	OEFP/FP	OEFN/FN	FP/TotF
Flat_Low	65.4 [37.1 76.8]	81.2 [66.1 88.5]	2.06 [0.94 6.67]	81.0 [60.0 90.1]	72.0 [50.2 86.3]	96.4 [88.9 99.4]	77.3 [58.8 88.3]
Flat_High	65.9 [37.2 77.1]	80.9 [64.8 89.1]	1.99 [0.96 6.51]	81.0 [60.6 90.0]	72.4 [50.1 86.7]	96.4 [89.5 99.4]	78.2 [60.1 88.9]
Slope	65.4 [36.7 76.7]	81.1 [64.2 88.8]	2.01 [0.96 6.50]	81.0 [59.6 90.3]	72.1 [48.5 86.5]	96.1 [88.1 99.3]	77.8 [57.5 88.3]

Table 7.3: Segmentation assessment measures when evaluating the influence of plateauing stages on the longitudinal framework. By contrast to Flat_High and Flat_Low, Slope refers to a pattern without plateauing values. Results are given under the format median [IQR] and are obtained across all subjects and common time points. Definitions of the assessment measures are given in Section 5.2.

rected difference lesion segmentations generated for each of the three compared methods. The corresponding results are summarised in Table 7.4. The pairwise statistical significance for the differences across methods for the DSC, TPR and AvDist were obtained with a paired Wilcoxon test. An increased performance in the order Cross $<$ Cross+ $<$ Long was observed and all tests were significant with p-value $<$0.0001 except for the average distance between the Long and Cross+ (p=0.13). The comparison of the DSC across methods for the different progression patterns is presented in Figure 7.7 illustrating the larger variances in assessment measures for the cross-sectional

		DSC	TPR	AvDist	OE/TotF	OEFP/FP	OEFN/FN	FP/TotF
Linear_500	**Cross**	18.1 [2.0 47.2]	15.5 [1.38 38.8]	11.38 [3.0 22.4]	55.5 [18.2 84.3]	38.2 [12.1 65.4]	65.2 [17.1 92.1]	9.5 [2.1 31.6]
	Cross+	61.0 [18.8 76.2]	73.6 [34.3 86.0]	2.33 [0.97 12.3]	79.6 [43.7 90.3]	70.5 [37.8 87.2]	92.9 [68.3 99.0]	65.3 [39.7 83.8]
	Long	64.9 [28.1 77.1]	83.3 [65.9 90.6]	2.07 [1.00 9.87]	80.5 [48.3 90.4]	72.1 [40.7 86.8]	96.6 [88.1 100]	82.7 [59.8 91.7]
Linear_750	**Cross**	26.9 [2.9 54.0]	19.6 [2.0 43.0]	6.44 [2.51 21.31]	61.2 [21.3 86.9]	42.2 [18.6 76.6]	66.1 [21.6 89.9]	8.8 [2.2 24.8]
	Cross+	66.4 [32.6 77.0]	72.5 [45.2 82.9]	2.00 [0.91 6.5]	80.5 [56.8 89.6]	69.2 [45.2 86.8]	87.8 [63.1 96.0]	60.6 [38.8 77.4]
	Long	66.3 [35.2 77.2]	80.2 [56.9 88.5]	2.00 [0.86 6.06]	80.2 [57.4 90.6]	70.3 [45.7 87.0]	94.0 [81.0 98.0]	78.9 [56.4 87.9]
NonLinear_5	**Cross**	14.2 [0.8 43.2]	9.1 [0.4 30.3]	11.19 [3.63 22.68]	55.1 [18.8 84.6]	31.3 [11.1 59.7]	59.5 [17.9 89.6]	6.1 [1.5 16.2]
	Cross+	63.8 [32.3 76.4]	73.0 [39.0 82.9]	1.98 [0.92 7.4]	81.6 [57.8 90.3]	69.9 [44.7 86.9]	92.8 [76.3 97.7]	57.0 [28.5 77.8]
	Long	66.1 [40.5 76.0]	81.0 [64.7 88.4]	1.90 [0.96 5.51]	82.4 [67.3 90.4]	74.1 [54.4 87.0]	97.4 [91.9 99.6]	77.1 [55.7 86.9]
NonLinear_15	**Cross**	21.0 [4.5 46.2]	14.3 [2.5 32.0]	8.64 [3.30 18.86]	59.0 [25.3 84.0]	41.2 [16.7 69.0]	61.8 [24.6 89.5]	7.8 [2.2 20.0]
	Cross+	59.8 [33.1 73.9]	70.0 [40.1 81.2]	2.34 [1.00 6.54]	80.4 [58.2 89.4]	69.9 [40.5 84.0]	91.1 [78.5 96.7]	56.0 [30.0 76.3]
	Long	64.4 [41.4 76.5]	78.6 [64.5 86.3]	2.02 [0.98 6.28]	79.1 [60.4 89.9]	70.4 [51.0 85.1]	95.2 [87.8 98.6]	75.2 [57.7 85.3]

Table 7.4: Segmentation assessment comparison for the three compared methods across all subjects and time points for all non plateauing patterns, subjects and time points. Results are given under the form median [IQR]. Definitions of the assessment measures are given in Section 5.2.

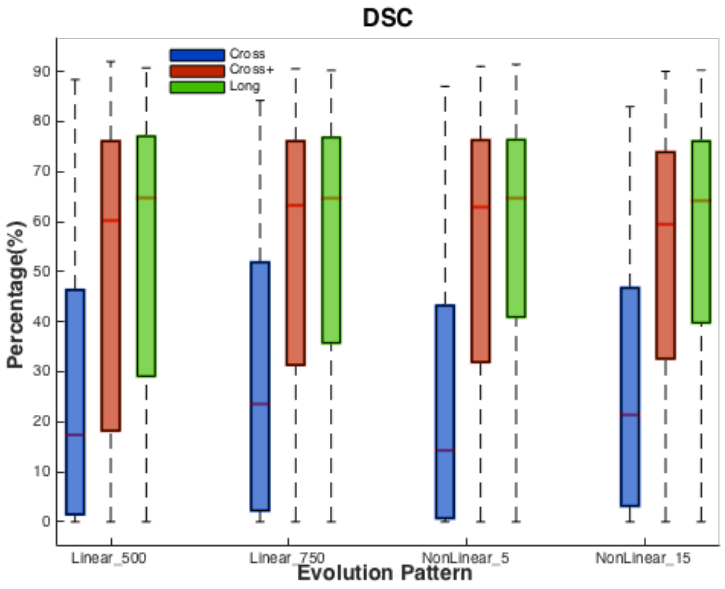


Figure 7.7: Comparison of the DSC distributions between the three methods across the different evolution patterns for the non-plateauing cases.

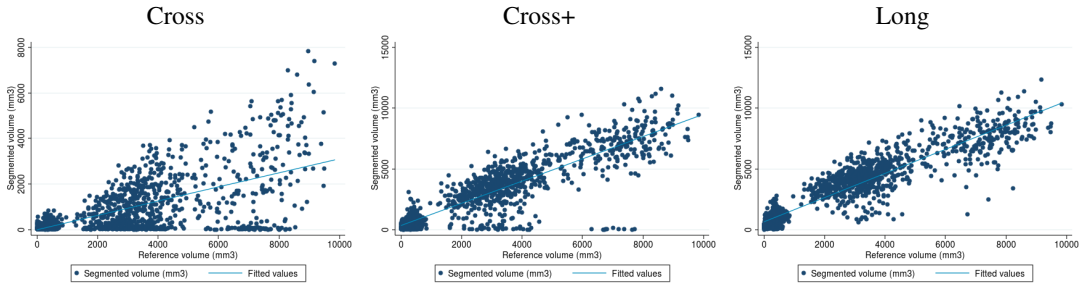


Figure 7.8: Regression plots of the segmented volume against the reference volume both corrected for the baseline segmentation intersection. Regression for the two cross-sectional methods (Cross and Cross+) and the longitudinal (Long) one are presented.

methods compared to the proposed longitudinal version.

Additionally, the robustness was tested through a linear regression of the ground truth volumes against the segmented volumes for all time points and subjects of the non-plateauing evolution patterns considering all time points independently. Figure 7.8 displays the regression plots while Table 7.5 summarises the regression parameters.

	Cross	Cross+	Long
Slope	0.31 [0.25 0.38]	0.92 [0.82 1.02]	0.99 [0.91 1.07]
Const	-5 [-115 106]	324 [142 507]	662 [480 845]
R^2	0.41	0.79	0.86

Table 7.5: Coefficients of the regression between segmented volume and reference volume for the three segmentation methods compared. The constant value is given in mm^3.

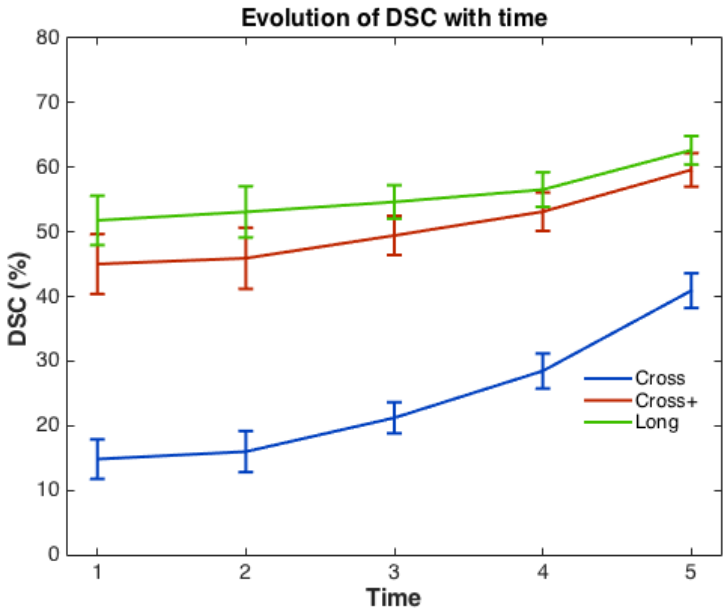


Figure 7.9: Mean DSC evolution in time with 95 % CI across the three evaluated methods for all subjects and non plateauing evolution patterns. Early time points correspond to smoother lesions and lower overall load.

Lastly, the longitudinal evolution of the DSC, corrected for the agreement volume at baseline, was compared across methods using a linear mixed model. Significant difference was observed in the slopes with a quicker decrease in DSC for the cross-sectional methods (p-value < 0.0001 and p-value $=0.007$ for Cross and Cross+ respectively). The corresponding plots of the predicted mean DSC are presented in Figure 7.9. This illustrates, above the effect of the lesion volume, the impact of the smoothness of the lesions on the quality of the segmentation.

7.3.4.3 Simulation of treatment effect

For the 51 concerned cases, the slopes of WMH volume change were assessed using linear mixed models to compare the fast evolution (linear change of 750 mm^3 per step) and the combined evolution (initial step of 750 mm^3 change followed by two steps of 500 mm^3 change) after the simulated treatment effect. The results of the estimation is presented in Table 7.6 exposing notably the lower measurement variance observed for the longitudinal framework compared to the cross-sectional methods.

7.3.5 Discussion

In this set up, the main strength of the simulator is also to some degree its main weakness. The realism of the simulator comes from the fact that real clinical data is used

		Cross	Cross+	Long
Treatment	**Mean**	439	532	370
	CI	[282 596]	[383 681]	[269 471]
No treatment	**Mean**	786	782	627
	CI	[597 975]	[629 935]	[525 729]
Statistics	**p-value**	0.006	0.022	0.004

Table 7.6: Slope estimation after evolution bifurcation for the three methods. Mean and confidence intervals (CI) are given for the two slopes.

both for the receiving and the donating set. However, the linear probabilistic intensity weighting using both the receiving set signal and the sampled lesion intensities may lead to some values incompatible with those expected according to the lesion ground truth. On one side, if a voxel is slightly hyperintense in the receiving image, it may appear truly hyperintense in the synthetic image even though the probabilistic weight would on its own not allow for such a classification. In turn, if lesion intensities with a relatively high probabilistic weight are to be combined with low intensities on the receiving set, due for example to the presence of iron, the resulting synthetic intensity may not satisfy the conditions necessary to be considered as lesion. Instead of drawing samples from typical distributions only for the lesion intensities, a possible alternative would be to also simulate healthy tissue distributions. In lieu of a probabilistic linear combination of intensities, another solution would be to derive the Mahalanobis distance associated to a given lesion probability, calculate the corresponding intensity based on existing inlier models and add centered Gaussian white noise afterwards. Moreover, other methods of intensity interpolation could be implemented and their impact studied. Lastly, it must be highlighted that the generalisability of the results obtained through the simulator is highly dependent on the donating lesion maps and receiving sets. Therefore, prior to simulation, a careful choice of the relevant dataset is strongly recommended.

Here, the simulation is applied to existing clinical data so that the realism of the simulated images is high. It must however be noted that the lesion simulator, although tested with different lesion loads, is based on typical age-related lesion distribution patterns as observed in the ADNI dataset that by design gathers subjects with limited cardiovascular risk factors (Maximum Hachinski score of 4). Increasing the variability of the lesion maps as well as the anatomical shape of change used for the simulator would

be of further interest. Since volumetric brain tissue changes occur also in ageing, simulation of atrophy patterns combined with lesion increase would further improve the realism of the simulator. With respect to the longitudinal framework, this simulator allowed the evaluation of the impact of evolution patterns on the segmentation performance as well as the impact of periods without change on the progression detection. From the assessment of the longitudinal framework across evolution patterns, subjects with smaller variations in WMH load led to slightly better overall segmentation results although no statistical difference was observed.

Investigating longitudinal bias through volume plateauing, the stability in the results highlights the ability of the longitudinal framework to detect change even if periods of no-change are included. This stability is crucial in processes for which subtle and irregular progressions are observed, such as multiple sclerosis. In terms of error, most of the erroneous classifications appear at the border of the lesions and very few lesions were completely undetected as illustrated by the high proportion of outline error. The comparison between the longitudinal and the cross-sectional versions of BaMoS underlines the improved robustness of the proposed framework with higher performance and lower variance in the results. Although a positive constant bias was observed with respect to the expected volumes, the strong correlation observed between segmented and reference volumes (R^2=0.86) makes the detection of change trustworthy. The ability to detect longitudinal rate of change was further exemplified in the simulations of treatment effect. In this case, the difference observed between evolutions is similar to the simulated ground truth difference. A decreased variance reflects higher measurement robustness compared to both cross-sectional methods and would translate into a lower required sample size in the context of a clinical trial.

7.4 Clinical application

The longitudinal framework was used to explore the evolution of WMH in the ageing population with respect to genetic status and especially APOE. While no direct quantitative validation is possible in this case, comparison with findings reported in the literature can highlight the relevance of the proposed method.

7.4.1 APOE and WMH

The APOE gene, located on the 19^{th} chromosome is present in the human population under the three main isoforms APOE $\varepsilon 2$, APOE $\varepsilon 3$ and APOE $\varepsilon 4$. Their expression results in different forms of the apolipoprotein E (apoE), plasma cholesterol transport protein highly implicated in the regulation of lipid transfer and lipolysis. The different effects of the isoforms in terms of neurodegeneration do not seem to be related to the differences occurring in terms of lipid metabolism. In terms of prevalence of the different alleles, APOE $\varepsilon 3$ is the more common before APOE $\varepsilon 4$ and APOE $\varepsilon 2$. If the liver is the main producer of apoE in the rest of the body, in the brain, astrocytes and to a lesser extent glia are the privileged production location [263].

The version $\varepsilon 4$ is widely recognised as a major risk factor for the incidence of AD and a decrease in the age of onset [264] and is thought to have a deleterious impact on the cerebral vasculature [265]. If the presence of one allele $\varepsilon 4$ increases the risk of developing AD 3-fold, this value is up to 12-fold in case of homozygosity. Although it affects the onset of the disease, impact on the disease progression is less certain. Required for the sustainment of neuronal plasticity, its activity in the repair of the nervous system appears to be lower in the case of the apoE4 isoform [266] and apoE is suggested to be implicated in myelin repair [16]. Neurotoxicity of cleaved forms of apoE have also been mentioned.

With respect to the Aβ pathological pathway, apoE isoforms have been shown to contribute differently to the fibrillisation of Aβ oligomers and to Aβ clearance with increased fibrillisation and impaired clearance in the case of the isoform apoE4 and contributes to an increased Aβ deposition both in the parenchyma and in the blood vessels [265, 266].

If the synergetic circle linking WMH, Aβ and amyloid deposition is widely documented [264, 265, 267, 268], other biological hypotheses have been entertained to explain the relationships between APOE status and WMH apart from amyloid pathology as reviewed by Tai et al. [269]. APOE $\varepsilon 4$ allele has indeed been linked with the permeability of the BBB and with a decrease in the tight junctions of the blood vessels' endothelium [270]. Presenting affected pericytes that contribute to the BBB effectiveness, in AD, APOE $\varepsilon 4$ carriers have been shown to display a higher BBB permeability [263]. An increased permeability could in turn lead to neuroinflammatory processes,

which may contribute to the development of WMH. Besides, the dose-dependent effect of the number of $\varepsilon 4$ alleles could be linked to the hypothesis of the protective effect of allele $\varepsilon 3$ on the BBB with respect to neuroinflammation [271]. Furthermore, APOE $\varepsilon 4$ has been associated with a decrease in glucose uptake, thus leading to deprived regions more vulnerable to ischaemia [272]. The homozygous $\varepsilon 44$ presentation of the APOE genotype has been associated with a reduction in capillary surface [273] in AD; this reduction would directly affect the blood supply in white matter, thereby promoting the development of WMH lesions. Further damage to the blood vessels associating for instance APOE $\varepsilon 4$ with microbleeds [274] and coronary heart disease [275] or stroke [276] could be lastly related to the exacerbated deleterious effects of vascular risk factors on WM in APOE $\varepsilon 4$ carriers [277, 278].

Considering the direct or secondary impact of APOE on the brain vasculature, associations have been investigated between age-related WM damage and APOE genetic status. Although meta-analyses tend to find a relation between APOE $\varepsilon 4$ and WMH burden [279], a few studies challenge this relationship [280, 281] and the question of an early effect on the white matter microstructure is controversial [282, 283].

7.4.2 Data and experiments

Publicly available, with T1 and FLAIR images, acquired at multiple time points for many subjects along with genetic samples, blood and CSF samples and neuropsychological evaluations, the ADNI (Alzheimer's Disease Neuroimaging Initiative) database arose as the database of choice to assess this longitudinal framework in a clinical setting.

At their initial visit, following clinical and neuropsychological assessment, each subject was given one of four diagnoses: Normal control (NC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) or probable AD. Subjects with a Hachinski score for cerebrovascular disease higher than 4 were excluded from the study thereby limiting the range of WMH load at baseline.

Subjects from the ADNI database were selected on criteria based both on scan relevance and genetic status. Those with an APOE $\varepsilon 2$ allele were excluded as it is thought to have a protective effect against over-production of A$\beta 42$ [284]). Only subjects for whom at least four imaging time points with serial FLAIR scans were available were

selected. In order to avoid introducing any preprocessing bias into the analysis, only subjects with T1 scans preprocessed with N3 histogram sharpening, and corrected for B1 bias field and gradient non linearity were used. Subjects were further excluded if they had poor quality imaging.

The genotyping for APOE was performed at screening visit using DNA extracted by Cogenic using a 3 mL aliquot of EDTA blood while the level of CSF Aβ is obtained using the xMAP Luminex platform and the Innogenetics/Fujirebio AlzBio3 immunoassay kits. Further details on the diagnostic procedure, scanning and imaging protocols as well as genotyping and Aβ measurement can be found at `http://www.adni-info.org/scientists/ADNIStudyProcedures.aspx`.

The longitudinal BaMoS framework was applied to the different sets of images while the total intracranial volume (TIV) is automatically obtained from the average image using the previously mentioned GIF pipeline [198].

7.4.3 Statistical analysis

All the statistical analyses were performed using Stata 12 v1. Due to the skewness of the WMH volumes, they were log-transformed. Cross-sectional analysis of the log-transformed WMH volumes at baseline was performed using linear regression models including age, TIV and sex as covariates in all models. Four models were fitted: these included diagnosis status (Model ASTD), genetic status (Model ASTG), diagnosis status and genetic status (Model ASTGD), and genetic status and CSF Aβ concentration (Model ASTGAβ) as predictor variables. Joint F-tests were used to assess differences between groups after adjustment for covariates. Fitted group specific means, standardised to the mean levels of covariates in the sample as a whole (with 95% confidence intervals), were also computed and back-transformed.

Longitudinal changes in WMH volume were assessed using linear mixed models, with random intercepts and slopes, for the repeated measures. Linear mixed models, provided that they are properly specified, appropriately allow for the non-independence of repeated measures from the same subject [285]. The dependent variable in all models was the log-transformed volume of WMH with time from initial measurement treated as both a fixed and random effect (thereby allowing slopes to differ between subjects). Other fixed effects were group terms (diagnosis and/or APOE status) and group-time

interaction terms (thereby allowing mean rates of change to differ between groups) and age, TIV and sex and their interactions with time. One model (Model ASTG) investigated differences in slope between the APOE groups, a second (Model ASTD) differences between the diagnostic groups and a third (Model ASTGD) differences according to both of these factors simultaneously. A fourth model (Model ASTGAβ) investigated APOE status and CSF Aβ concentration as predictors of rates of decline. Joint Wald tests were used to compare rates of change between groups after adjustment for covariates. Fitted group specific mean rates of change, standardised to the mean levels of covariates in the sample as a whole (with 95% confidence intervals), were computed and back-transformed for each model.

7.4.4 Results

7.4.4.1 Demographic results

The inclusion criteria led to the selection of 300 subjects that collectively underwent 1430 scans. The number of acquired time points varied from 4 to 7 (mean 4.77 SD 0.73) and the total length of time since initial assessment varied from 11 to 52 months (mean 24.0, SD 9.4). The demographics of the included sample are presented in Table 7.7 by genetic status and diagnostic group. As expected, a decrease in CSF amyloid level was observed with an increasing diagnosis severity and an increasing number of APOE ε4 allele. Although age was comparable across APOE status, EMCI and LMCI were younger than NC and AD.

7.4.4.2 Cross-sectional associations of WMH

The baseline data are summarised in Table 7.8. There was evidence (p = 0.030, Model ASTD) that the volume of WMH differed between the diagnosis severity groups, the difference being mostly driven by the low volumes observed in NC compared to the three other groups. Although the mean in the AD group was slightly lower than that in the EMCI and LMCI groups, these differences were not statistically significant and the 95% confidence interval for AD group mean was wide, reflecting the fact that this group contains the fewest subjects. Similar results were seen when the differences between the diagnostic groups were adjusted for APOE genetic status (p-value from joint test of differences = 0.066, Model ASTGD). Considering the Aβ level as a continuous marker also provided evidence of an association (p<0.0005) with (Model ASTGAβ) or

	Diagnosis	APOE status 33	43	44	**Global**
Number Total [Female Aβ]	NC	56 [27 53]	18 [9 15]	3 [0 3]	**77 [36 71]**
	EMCI	76 [37 73]	52 [21 50]	4 [0 3]	**132 [58 126]**
	LMCI	27 [12 26]	24 [17 23]	11 [5 11]	**62 [34 60]**
	AD	9 [3 9]	14 [5 14]	6 [1 6]	**29 [9 29]**
	Global	**164 [79 161]**	**108 [52 102]**	**24 [6 23]**	**300 [137 286]**
Study duration (months) mean (SD)	NC	23.1 (7.4)	21.9 (4.9)	25.3 (0.6)	**22.9 (6.8)**
	EMCI	26.1 (10.4)	27.4 (12.1)	30.8 (12.2)	**26.7 (11.1)**
	LMCI	23.9 (5.6)	23.4 (8.1)	19.1 (6.0)	**22.8 (6.9)**
	AD	14.6 (4.1)	15.8 (5.9)	20.5 (6.2)	**16.4 (5.7)**
	Global	**24.1 (8.9)**	**24.1 (10.4)**	**22.2 (7.9)**	**24.0 (9.4)**
WMH (mL) median [IQR]	NC	1.67 [0.76 3.54]	1.90 [1.33 5.72]	2.78 [2.15 17.46]	**1.85 [0.78 4.33]**
	EMCI	1.91 [0.88 5.63]	2.75 [0.56 6.08]	0.92 [0.83 12.9]	**2.08 [0.70 5.84]**
	LMCI	2.22 [1.15 4.45]	2.02 [0.60 5.02]	3.37 [1.78 9.10]	**2.33 [1.08 5.66]**
	AD	8.31 [2.06 10.51]	3.37 [1.44 9.39]	0.98 [0.49 4.08]	**3.52 [1.31 8.34]**
	Global	**2.06 [0.88 5.42]**	**2.48 [0.68 6.08]**	**2.93 [0.96 5.64]**	**2.15 [0.88 5.74]**
Age mean (SD)	NC	74.2 (5.8)	73.0 (8.0)	78.1 (9.7)	**74.1 (6.5)**
	EMCI	71.6 (7.2)	69.7 (7.2)	68.4 (5.1)	**70.8 (7.2)**
	LMCI	71.4 (7.5)	71.4 (8.1)	70.6 (7.2)	**71.3 (7.6)**
	AD	78.9 (6.0)	74.6 (7.3)	70.1 (7.2)	**75.0 (7.3)**
	Global	**72.9 (7.0)**	**71.3 (7.7)**	**71.1 (7.3)**	**72.1 (7.3)**
TIV (mL) mean (SD)	NC	1530 (158)	1543 (152)	1680 (458)	**1539 (164)**
	EMCI	1564 (123)	1564 (165)	1720 (456)	**156.9 (141)**
	LMCI	1553 (158)	1516 (201)	1492 (127)	**152.8 (171)**
	AD	1568 (178)	1557 (173)	1537 (131)	**1556 (162)**
	Global	**1552 (143)**	**1549 (171)**	**1564 (141)**	**1552 (153)**
Aβ (ng/L) mean (SD)	NC	200.6 (44.4)	172.9 (41-8)	111.0 (42.3)	**190.9 (47.0)**
	EMCI	197.7 (46.8)	170.4 (47.7)	121.1 (14.0)	**185.1 (49.3)**
	LMCI	196.4 (44.9)	138.1 (32.5)	113.9 (31.4)	**158.9 (50.8)**
	AD	156.4 (52.2)	143.2 (41.2)	101.6 (10.4)	**138.7 (44.6)**
	Global	**196.1 (46.3)**	**159.8 (45.0)**	**111.3 (26.4)**	**176.1 (51.6)**

Acronyms expansion: TIV - Total Intracranial Volume ; NC - Normal Control ; EMCI - Early Mild Cognitive Impairment ; LMCI - Late Mild Cognitive Impairment ; AD - Alzheimer's Disease ; Aβ - Aβ CSF level; WMH - White Matter Hyperintensities ; IQR - InterQuartile Range; SD - standard deviation

Table 7.7: Demographic data of the studied sample by APOE and diagnostic status.

without adjustment for APOE. Across APOE status, an increase in WMH volume with the number of $\varepsilon 4$ alleles was observed, although this was not statistically significant when APOE was considered with or without adjustment for diagnostic group or Aβ. In all these models, the impact of age and TIV were significant (p<0.001) and the gender difference was borderline significant for Model ASTG (p=0.094) such that females tended to have more WMH.

The baseline data are summarised in Table 7.8.

7.4.4.3 Models of longitudinal WMH volume change

Table 7.9 summarises the results for the longitudinal assessment of the WMH rate of change. Evolution rates are presented as adjusted mean values of percentage change in volume per year.

There was strong evidence (p=0.009) that rates of change differed between the

		APOE			Diagnosis			
		33	**43**	**44**	**NC**	**EMCI**	**LMCI**	**AD**
Model	Number (Aβ)	164 (161)	108 (102)	24 (23)	77 (71)	132 (126)	62 (60)	29 (29)
ASTG	Volume	1.96	2.26	3.14				
	CI	[1.65 2.32]	[1.83 2.79]	[2.00 4.92]				
	Overall p		0.13					
	Pairwise		/					
ASTD	Volume				1.58	2.27	2.72	2.18
	CI				[1.23 2.03]	[1.88 2.75]	[2.07 3.59]	[1.45 3.27]
	Overall p					0.030		
	Pairwise					NC vs EMCI * - NC vs LMCI **		
ASTGD	Volume	2.01	2.20	2.96	1.62	2.32	2.61	2.05
	CI	[1.69 2.38]	[1.78 2.72]	[1.87 4.71]	[1.26 2.08]	[1.91 2.81]	[1.97 3.46]	[1.3 3.10]
	Overall p		0.30			0.066		
	Pairwise		/			NC vs EMCI * - NC vs LMCI *		
ASTGAβ	Volume	2.18	2.05	2.21				
	CI	[1.82 2.63]	[1.64 2.56]	[1.34 3.65]				
	Overall p		0.90					
	Pairwise		/					

Table 7.8: Baseline models: effects of covariates on differences in WMH volumes across diagnostic groups and APOE genotypes when adjusting for age sex TIV. Results are presented under their back-transformed format.

APOE groups with the largest mean rate in the APOE 44 group and the smallest in the APOE 33 group. The mean rate in the APOE 44 group (15.5%/year) was significantly higher than in each of the other two groups (8.7%/year in the 43 group and 5.7%/year in the 33 group). The mean rates in these latter two groups did not differ significantly from each other. Adjustment for either diagnostic group (Model ASTGD) or Aβ42 level (Model ASTGAβ) attenuated the differences between the groups without altering the overall pattern. In the latter model the dependency of the rate of change on the CSF Aβ42 level was statistically significant (p=0.003) and the observed trend lost significance (p=0.39).

The variability observed within diagnostic did not allow for any significant pattern

		APOE			Diagnosis			
		33	**43**	**44**	**NC**	**EMCI**	**LMCI**	**AD**
Model	Number (Aβ)	164 (161)	108 (102)	24 (23)	77 (71)	132 (126)	62 (60)	29 (29)
ASTG	% change	5.68	8.68	15.53				
	CI	[3.56 7.84]	[5.95 11.48]	[9.09 22.34]				
	Overall p		0.009					
	Pairwise		33 vs 44 **					
ASTD	% change				6.99	6.83	9.38	8.51
	CI				[3.63 10.46]	[4.48 9.23]	[5.57 13.32]	[1.44 16.06]
	Overall p					0.71		
	Pairwise					/		
ASTGD	% change	5.69	8.71	15.34	7.60	7.18	8.70	6.25
	CI	[3.50 7.92]	[5.94 11.55]	[8.75 22.34]	[4.22 11.09]	[4.85 9.58]	[4.94 12.29]	[-0.76 13.75]
	Overall p		0.015			0.90		
	Pairwise		33 vs 44 **			NC - EMCI * NC - LMCI *		
ASTGAβ	% change	6.81	7.82	12.18				
	CI	[4.47 9.21]	[4.96 10.76]	[5.32 19.49]				
	Overall p		0.39					
	Pairwise		/					

Table 7.9: Longitudinal models: effect of baseline predictors on differences in WMH volume change when adjusting for age sex and TIV. Adjusted means of percentage of change are presented along with the confidence intervals.

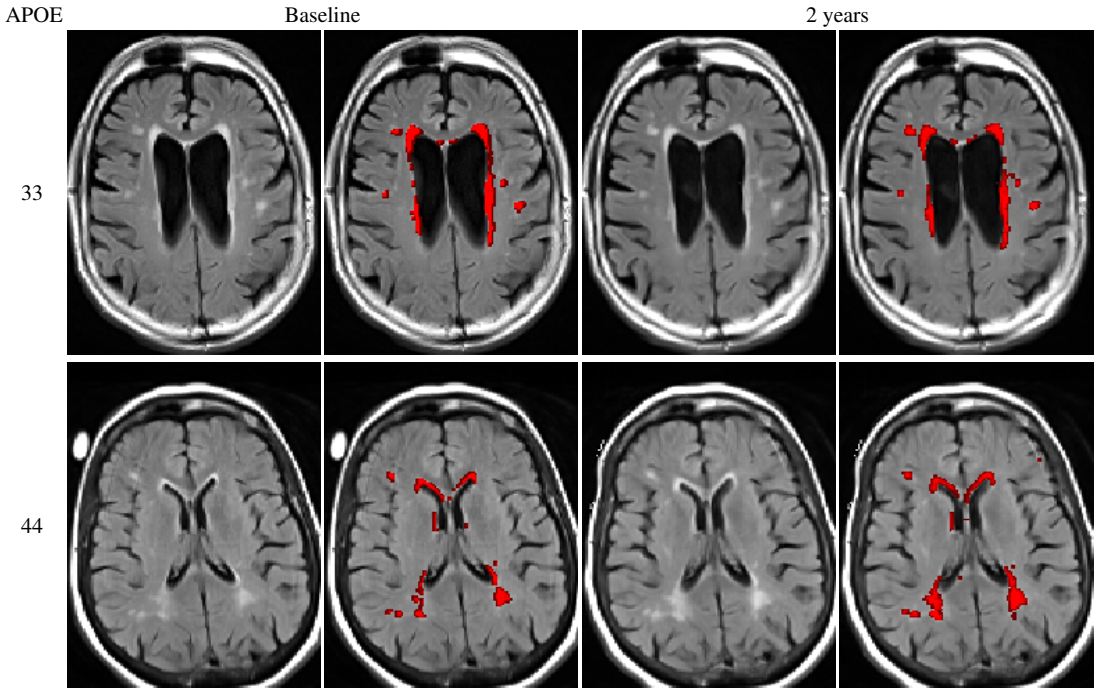


Figure 7.10: Longitudinal segmentation for two LMCI subjects with status 33 (first row) and 44 (second row) with same lesion load at baseline (left). Note the faster rate of accumulation for the homozygous APOE $\varepsilon 4$ subject (right).

to arise across diagnostics. Again, the CI for AD was wide and its higher extremity higher than for the other groups.

Figure 7.10 presents the evolution in WMH for two LMCI subjects with same load at baseline one with status 33 and the other with status 44 at first scan and after 2 years.

7.4.5 Discussion

This longitudinal study shows a strong association between the APOE status and the rate of WMH volume accumulation. An increased rate of change was observed for the homozygous 44 compared to 33 and 43 carriers in a cohort with relatively low vascular burden at baseline. The subjects in the study ranged from healthy controls through people with mild cognitive impairment to Alzheimer's Disease patients. These findings were observed independently of diagnosis and the same trend although non significant existed when correcting for CSF Aβ levels.

In the ADNI study, exclusion criteria were designed to limit the amount of vascular disease using a threshold of 4 on the Hachinski score. Therefore, results obtained from that sample cannot be immediately generalised to a wider population. For in-

stance, the percentage change of 15.5 % observed for the APOE 44 group might be less likely to occur at higher initial loads. Compared to a general population, this cohort appears enriched in subjects carrying APOE $\varepsilon 4$ alleles, which may have helped in extracting relevant patterns and allowed the distinction between heterozygous 43 and homozygous 44. The bias towards a vascular-risk free population may further explain why no statistically significant differences at baseline were observed in this sample with respect to APOE status. Additionally, the differences observed in terms of study length for the LMCI and AD groups may have led to an underestimation of the observed pattern differences. In the case of the AD population, an additional selection bias might be added since subjects cumulating higher loads of WMH and AD pathology may drop out quicker than those with little WMH; in AD, the selection criteria of four time points may be more difficult to achieve. Despite these potential limitations, a slight trend for increased volume and rates of change with diagnosis severity was still observed.

Initial volumes of WMH have been reported to be positively related to rates of change in lesion load in both normal [128, 286] and demented populations. The possibility of a non-linear accrual of WMH naturally raises the question of how best to model accelerating biological processes in statistical analyses. Since volumes are bounded by the brain size, logarithmic models are imperfect and results are difficult to compare with observations published based on the raw data. However, in the log-transformed framework, a relationship between initial raw measurement and absolute measure of change is inherently allowed for since the log-transformed analysis models changes on a relative scale.

There was evidence that different patterns of WMH evolution were related to APOE status consistent with results reported by Godin et al. [287] for normal ageing. In this study, a much higher rate of increase in subjects homozygous for APOE $\varepsilon 4$ compared to heterozygous or non-carriers was reported. The finding of no statistically significant difference in rate of change between heterozygous 43 and non carriers may be taken as a possible explanation for controversies regarding the link of APOE $\varepsilon 4$ with WMH evolution since in some studies the population was simply dichotomised into carriers and non-carriers [288]. Yet, the different association strengths observed with the covariates for the 43 and 33 carriers may reflect potential differences in the pathological process between these two groups and would deserve further investiga-

tion.

The association between amyloid pathology, APOE and WMH probably reflects the link between Aβ, APOE ε4 and cerebral amyloid angiopathy (CAA) [39,264,269]. Since a similar trend was observed after adjustment for diagnosis group or Aβ level, the additional effects of APOE on the vasculature, independently of amyloid, play certainly a role in accounting for the much faster increase in WMH for subjects homozygous in APOE ε4.

The possible joint effects and interaction of vascular risk factors such as hypertension and APOE ε4 on the blood vessels is therefore in need of further investigation with respect to the development of WMH. So far, cross-sectional studies have reported controversial findings regarding the relationship between APOE ε4 and WMH. This is further illustrated here by the absence of difference across genetic status at baseline while significantly different rates of change are observed. This finding strongly supports the need to better understand the time course and development pattern of pathological processes. Limitations of cross-sectional studies may be partially overcome by the application of longitudinal models.

This study could, however, be expanded in a number of ways. Due to the complexity of the possible biophysiological interactions between amyloid compounds and apolipoprotein, any statistical conclusion should be treated cautiously. As such, CSF level of Aβ-42 used as a surrogate disease marker for AD pathology [289] is also known to be related to WMH [290]. Aβ-40 measures, more associated to vascular deposition, could prove interesting to better understand the pathological process. Furthermore, only global WMH loads were considered here, but regional assessment of the lesion growth and differences in pattern across regions could improve the understanding of APOE and AD pathology combined effects. Combining the systematic description of lesion distribution detailed in Chapter 6 and the longitudinal method developed here may highlight more specific patterns of evolution. Besides, allele ε2 carriers were excluded from this study. A possible link between the allele ε2 and WMH has however been reported [291] but the very low prevalence of this allele in the population makes any investigation difficult.

In conclusion, this clinical application of the longitudinal framework has shown the influence of APOE ε4 on the rate of WMH accrual over and above diagnosis status.

Carriage of a single $\varepsilon 4$ allele gave a non- significant additional increase of 3% per year, whereas homozygous carriers had a significant additional increase of 10% per year compared with $\varepsilon 3$ homozygotes. APOE $\varepsilon 4$, especially when in the homozygous form, is an important independent factor in the progression of WMH.

7.5 Discussion

In this framework, the average image on which a generative data model is then derived enables the reduction of measurement noise and accounts for the within-subject correlation. Accounting for such correlation instead of simply applying cross-sectional methods for quantitative assessment has indeed been shown to reduce the measurement variability [259]. Furthermore, the creation of an average image in a midspace overcomes the problem of a bias towards a specific time-point as previously mentioned [251, 252]. Lastly, the use of the generative model to constrain the segmentation at each time point is in itself a guarantee that subtle changes can be accounted for while maintaining robustness.

Various points could nonetheless benefit from further investigation in order to better evaluate the robustness and validate the procedure. For instance, the arbitrary choice of the number of steps required to build the average image and the dependence on the registration parameters would be of interest. Moreover, since a polynomial fit of degree 2 may not be too greatly affected by the presence of outliers, investigating the differences observed when excluding obvious outliers in the fit or not may put some light into their impact on the average image creation.

Apart from the technical details of the framework, validation perspectives, especially taken with respect to clinical findings, must be considered with caution. As observed in the case of BaMoS, small lesion loads are generally more difficult to handle resulting in a higher uncertainty in the measurement. This tendency is strongly reduced when considering a full set of longitudinal data. Such observation may have a direct and important consequence when comparing longitudinal and cross-sectional method for clinical studies. The artificially lower loads observed at early time points of the evolution in the cross-sectional case can indeed result in a naturally higher slope of evolution than with the longitudinal system. Due to this bias, differences between groups could then reached different significance levels when using the longitudinal or

the cross-sectional framework. This naturally leads to the question of the reproducibility of clinical observations, for which the dependence on the bias of the employed automatic techniques should be acknowledged. At times when early evolution and development become the focus of state-of-the-art research, assessing confidently uncertainty and bias in numerous conditions of processing is becoming essential. Moreover, type, shape and direction of evolution may further affect the longitudinal assessment of lesion evolution. In fact, with partial volume effects at the border of the lesions, direction of lesion progression (isotropic or anisotropic) combined with the image resolution may lead to different change detection although with the same volumetric variation. Similarly to what has been reported in Section 5.2.2 with respect to the DSC, lesion size and shape may influence the detection ability, highlighting further the greater uncertainty inherent to lower lesion loads measurements. Shifts in WMH volume appear therefore insufficient to describe the longitudinal evolution of the pathology. Complementary information regarding direction of expansion, apparition of new lesions presented in the perspective of other modalities informative on WM tracts and vascularity would prove extremely valuable to continue improving longitudinal algorithms and understanding lesion evolution.

Chapter 8

Summary and future work

8.1 Summary and limitations

Unexpected signal can occur in medical images due to pathological conditions or mechanistic acquisition defects. These observations must be taken into account not only because they affect models and evaluation of normal aspects but also because they are assumed to bear a biological meaning relevant to the understanding of pathological conditions. The need for automated quantification of medical imaging observations stems from the requirements of robustness and generalisability of measurements used to better assess clinical hypotheses and establish pathophysiological pathways in large cohorts. The high variability inherent to pathological presentation in the ageing population strengthens the need for generic, robust and flexible solutions.

Chapter 4 introduced a generic data modelling framework that uses a hierarchical GMM to jointly model normal and unexpected observations. Instead of imposing a predefined number of Gaussian components, the model complexity is automatically derived based on a split and merge strategy making the model more representative of biological tissues. Decoupling the modelling of the data from the pathological quantification allows the proposed model to remain independent from the clinical application. As described in Chapter 5, the model can then be used to segment WMH. Heuristic rules, independent of the model, were designed to be applied as a post-processing step to fit the clinical description and avoid the inclusion of false positives.

The internal evaluation performed in Section 5.4 showed the robustness of the lesion segmentation framework with respect to variations in preprocessing, modality or acquisition protocol choices. Intrinsic to the lesion segmentation problem lies the ques-

tion of the distinction between normality and abnormality and probabilistic representations are therefore required to account for the continuity in the damage. Noticeably, it is in areas of greater uncertainty at the lesion border that most differences between segmentation results were observed.

The external validation based on comparison to manual segmentations showed that BaMoS was more robust compared to other available lesion segmentation algorithms. As detailed in Section 5.2, multiple means of assessing a segmentation with respect to a gold standard reference are needed in order to acquire a complete picture of the performance of an algorithm and overcome the limitations of each single evaluation measure. However, as stated in Section 5.2.4, due to the inconsistencies inherent to manual delineations, such measurements cannot be used as the only evaluation source. Thus, surrogate measurements, such as the consistency in the intensity characteristics of the resultant segmentation need to be included in the assessment. It is then the combination of all the measurements and the careful analysis of the source of discrepancies that can contribute to the improvement of automated methods.

In Chapter 6, local lesion statistics were estimated using a subject specific geometrical WM regional parcellation. A graphical representation of the regions was designed to reflect these local statistics. The proposed regional analysis was used to estimate the inherited component of vascular damage in twin pairs and to the prediction and characterisation of rating scales. The results of these two studies demonstrated that the proposed regional coordinate frame and associated local WMH statistics were able to accurately describe WM disease patterns beyond global statistics. An online training tool, aiming to standardise ratings and provide trainees a feedback on their local bias towards lesions was derived from this representation.

In Chapter 7, the cross-sectional GMM framework was then extended to longitudinal data. A longitudinal lesion simulator was created to validate this model, when compared to the existing cross-sectional framework and results showed an increased robustness and sensitivity when using the longitudinal framework. The proposed longitudinal method was used to explore the relationships between APOE status and WMH accumulation. Due to its high sensitivity and specificity, the proposed longitudinal method showed highly significant differences in WMH accumulation between genetic groups, with stronger associations than have been previously reported clinically.

8.2 Future work

In the course of this work, many choices, opportunities and decisions with regards to research avenues to close or to explore have been made. It is thus impossible to properly appreciate how this research would have progressed if these decisions had been different.

Given the pool of work developed during the last three years, different aspects could be further investigated. With respect to the BaMoS model, additional *a priori* information could be included to constrain the evolution and the range of possible mixture models. Alternatively a different set of constraints may be applied to the covariance matrices in order to account differently for the various modalities. The split and merge strategy could also be replaced with a model sampling scheme, which would require novel technical developments with regards to model sampling and model averaging.

The main application of BaMoS targeted the segmentation of WMH. Other post-processing rules may however enable the quantification of other neurological markers such as enlarged perivascular spaces or lacunes. Priors over mixture model configurations and parameters for specific applications/pathologies could help constrain the space of solutions, resulting in more biologically meaningful mixture classes and more relevant models. Also, if different priors were available for different pathologies, Bayesian model comparison could then be used to select the best model and possibly even classify the patient's pathology. This could be further enhanced by an analysis based on the model subclasses of the texture of the different lesions. Alternatively, BaMoS may be applied to other conditions such as tumor segmentation and its inherent ability to distinguish different types of outliers used to separate tumor tissue types.

Within the WMH application, other sources of data could be used to further explore the source of pathophysiological change by looking for instance at the location with respect to white matter tracts and blood supply using diffusion imaging and arterial spin labelling.

Given the general nature of the BaMoS model fit, a series of hand crafted hierarchical rules were used to select relevant elements after the model fit. Another approach could be to use machine learning techniques to extract relevant classes/features for the segmentation of specific pathologies. Such hybrid method of combining generative model outcome with machine learning techniques has recently been shown to be quite

successful in the context of tumor segmentation and stroke [292].

With the recent success of learning oriented methods and in particular deep-learning, one may wonder if those techniques may solve completely the problems of characterisation and segmentation of pathological data. The understanding of the boundary between normality and abnormality is however not clear cut. Additional knowledge and understanding may thus be reached by the combination of learning processes and model building either using learning results to constrain and enhance models or using the richness of information provided by models to better feed learning processes.

Overall, solutions for the quantification of medical images go hand in hand with the rise of new clinical questions and hypotheses, and the improvement of imaging techniques. Collateral findings derived from quantification may enable new biological hypotheses to be proposed. It is therefore crucial to go beyond the simple reproduction of manual tasks and make use of enriched and robust information to improve patient care.

Appendix

Acronyms and abbreviations

The following acronyms are ordered alphabetically and not by order of appearance.

AD	Alzheimer's disease.
ADF	Anisotropic Diffusion Filter.
ADNI	Alzheimer's Disease Neuroimaging Initiative
AvDist	Average distance
BBB	Blood Brain Barrier.
BaMoS	Bayesian Model Selection.
BaMoS-static	Version of BaMoS, with no evolution of the model.
BaMoS-No-Cov	Version of BaMoS, with no constraint over the covariance.
BaMoS-nc	Version of BaMoS, not corrected for FP.
BF	Bias field.
BG	Basal Ganglia.
BGIT	Basal Ganglia and Infratentorial region.
BiASM	Bilayered Anatomically constrained Split and Merge expectation maximisation algorithm
BIC	Bayesian Inference Criterion.
BM	Refers to the BrainMaps methods for mask creation.
CC	Connected component.
CG	Centre of gravity.
CI	Confidence interval.
CMB	Cerebral microbleeds.
Cross	Cross-sectional version of BaMoS with flat outlier priors.
Cross+	Cross-sectional version of BaMoS with adapted typicality map for outlier priors. SE stands for sensitivity enhanced.

CSF	Cerebrospinal fluid.
CSVD	Cerebral Small vessel disease.
DE	Detection error.
DSC	Dice Similarity Coefficient.
DTI	Diffusion Tensor Imaging.
DWI	Diffusion Weighted Imaging.
ECM	Expectation Conditional Maximisation.
EM	Expectation-Maximisation.
EMCI	Early Mild Cognitive Impairment.
EMS	Expectation Maximisation Segmentation tool.
EMS-c	Clinical version of EMS (with MRF as defined in Section 5.4)
EMS-d	Default version of EMS (with adaptive MRF).
E-step	Expectation step.
EPVS	Enlarged perivascular spaces.
FCM	Fuzzy-C Means
FFE	Fast Field Echo.
FLAIR	FLuid Attenuation Inversion Recovery.
FN	False negative.
FNR	False negative rate.
FP	False positive.
FP/TotF	Proportion of false positives among all errors.
FRONT	Frontal lobe.
FPR	False positive rate.
FWHM	Full Width at Half Maximum.
GIF	Geodesic Information Flow.
GM	Grey matter.
GMM	Gaussian mixture model.
GS	Gold standard.
GT	Ground Truth.
ICBM	Refers to the priors derived from the ICBM consortium template.
ICM	Iterative Conditional Mode.

IIH	Intensity inhomogeneity.
IQR	Inter-quartile range.
KLD	Kullback-Leibler Divergence.
LA	Leukoaraiosis.
LMCI	Late Mild Cognitive Impairment.
Long	Longitudinal version of BaMoS with adapted typicality map for outlier priors.
LST	Lesion Segmentation Tool.
LST-WML	Clinical version of LST (0.25) optimised for WMH.
LST-MS	Default version of LST (0.3) optimised for MS.
MAP-EM	Maximum a Posteriori Expectation-Maximisation.
MNI	Montreal Neurological Institute.
MRF	Markov Random Field.
MR(I)	Magnetic Resonance (Imaging).
MS	Multiple sclerosis.
M-step	Maximisation step.
NABT	Normal appearing brain tissue.
NB	Non-Brain.
NC	Normal Control.
OCC	Occipital Lobe.
OER	Outline error rate.
OEFP/FP	Proportion of false positives belonging to the outline error.
OEFN/FN	Proportion of false negatives belonging to the outline error.
OE/TotF	Proportion of errors belonging to the outline error.
PAR	Parietal lobe.
PD	Proton density weighted.
SD	Standard deviation.
SM	Split and merge.
SVD	Singular Value decomposition.
T1	T1-weighted.
T2	T2-weighted.
T2DB	Type 2 Diabetes.

TE	Echo time.
TEMP	Temporal lobe.
TI	Inversion time.
TIV	Total intracranial volume.
TLE	Trimmed-likelihood estimator.
TLL	Total lesion load.
TLL_{auto}	TLL obtained by an automated method.
TLL_{manual}	Manually segmented TLL.
TN	True negative.
TP	True positive.
TPR	True positive rate.
TR	Repetition time.
TSE	Turbo Spin Echo.
VD	Volume difference.
WM	White matter.
WMC	White matter change.
WMH	White matter hyperintensity.
WML	White matter lesion.

Mathematical notations

In this work, a non bold lower case symbol corresponds generally to a scalar. A bold lower case symbol refers to a vector while an upper case bold symbol corresponds to a set of vectors. A letter in calligraphic writing generally corresponds to a known function. Depending on the context, the notation f can either refer to a distribution density function or to a probability function.

Sub-, Superscripts and generalities

$^{(t)}$	Iteration t.
^{c}	Corrected for IIH.
$^{\tau}$	Time point of longitudinal series.
$^{\mathrm{GW}}$	Groupwise space in a longitudinal series
^{T}	Applied to a vector or a matrix, denotes the transposition operation.
$^{(d)}$	Taking the value d or Patho, T1, T2, denotes a specific channel.
$\lvert\,\rvert$	Applied on a matrix, denotes the determinant of the matrix.
$\mathbb{E}$	Applied on a random variable, denotes the expectation of this variable.
$\boldsymbol{e}_{\ell}$	The ℓ vector of the canonical basis.
$\widehat{\ }$	Denotes the optimised version of the parameters.
$\star$	Indicates a convolution operation.

Usual notations

$\delta(x)$	Dirac function
$\gamma(n,x)$	Incomplete gamma function
$\Gamma(n)$	Gamma function
λ	Denotes a Lagrange multiplier

Indexing and counters

n and N	Applied to the voxels of the image, with index n and total number N.
m and M	Applied to the IIH polynomial basic functions, with index m and total number M.
d and D	Applied to the number of image modalities, with index d and total number D.
l	Used for the Level 1 of the model hierarchy and assume the value I or O.
j and J	Used for the Level 2 of the hierarchy with index j, and J representing the total number of anatomical classes.
k and K_{l_j}	Used for the Level 3 of the hierarchy with index k, and K_{l_j} representing the number of Gaussian distributions used to model class l_j.
k and K	Used only in Chapter 3, K represents the number of components in the considered GMM indexed by k.

Sets and vectors

y Indexed by n and of size D, refers to the vector of normalized log-intensities.

Y Set of vectors $\mathbf{y}_n$, with n varying from 1 to N.

z Indexed by n, represents the unity vector of the canonical basis characterising labelling configuration for voxel n.

Z Set of vectors **z**, represents the full labelling configuration for the images (hidden data).

Density distributions and parameters

$\boldsymbol{\mu}$ Mean (vector) for a Gaussian distribution.

Λ Covariance matrix for a Gaussian distribution.

Ω Weighted covariance matrix in a GMM.

θ Set of Gaussian parameters $\{\boldsymbol{\mu}, \Lambda\}$, generally indexed by l_{j_k}.

Θ Parameters of the Gaussian components of a mixture indexed by l_j.

K Model under consideration, characterising the number of Gaussian components K_{l_j} per mixture l_j. In chapter 3 refers to the model built from K Gaussian components.

$\boldsymbol{\Xi_K}$ Denotes the complete set of parameters used for model **K**.

I and O Denotes the density distribution function for the inlier and outlier part of the model respectively.

Φ Density distribution for a mixture at Level 2 of the hierarchy, indexed by l_j.

$\mathscr{L}$ Notation adopted for the marginal log-likelihood.

$\mathscr{G}$ Notation adopted for a Gaussian density distribution.

$\mathscr{U}$ Notation adopted for a uniform distribution.

$\mathscr{M}$ Generic notation for a distribution at Level 3: can be either uniform ($\mathscr{U}$) or Gaussian (G).

$\mathscr{W}^{-1}$ Generic notation for the Inverse Wishart distribution

Bias Field correction

BF Refers to the bias field correction function

χ Indexed by m, corresponds to a IIH polynomial basis function.

$\boldsymbol{\chi}$ Refers to the matrix built from the value taken by the M basis functions at the N considered locations.

v Indexed by n, corresponds to the spatial location of voxel n.

$\tau_n d$ Introduced in the bias field correction process to account for the weight attributed to the observation at location n for channel d.

Υ Indexed by d, corresponds to the diagonal matrix built with the N τ_{nd} values corresponding to channel d.

$\bar{y}_{nd}$ Introduced in the bias field correction process.

$\mathbf{r}$ Indexed by d, corresponds to the vector of size N of residuals defined at each location for channel d as $y_{nd} - \bar{y}_{nd}$.

c_{md} Refers to the bias field linear coefficient applied to function ξ_m for channel d.

$\mathbf{c}_m$ Denotes the vector of bias field linear coefficients of size D applied to basis function ξ_m.

$\mathbf{c}^d$ Denotes the vector of bias field linear coefficients of size M relative to channel d.

$\mathbf{C}$ Denotes the set of vectors of bias field coefficients $\mathbf{c}_m$.

Mixing weights and atlases

8.2.0.1 Generic GMM - Chapter 3

w Indexed by l_{j_k}, denotes the mixing weight of the component l_{j_k} in mixture l_j.

$\mathbf{w}$ Indexed by l_j, represents the vector of mixing weights $w_{l_{j_k}}$ used to model mixture l_j.

$\mathbf{W}$ Denotes the set of all vectors $\mathbf{w}_{l_j}$ in the model.

a (b) When simply indexed by l_j (l), corresponds to the global mixing weight of class l_j (l) at Level 2 (Level 1). When also indexed by n, corresponds to a voxelwise *a priori* probability.

$\mathbf{a}$ ($\mathbf{b}$) In the context of statistical atlases, corresponds to the vector built from a_{nl_j} (b_{nl}), indexed by n. Otherwise, corresponds to the vector of global mixing weights.

$\mathbf{A}$ ($\mathbf{B}$) Set of vectors $\{\mathbf{a_1}, \cdots, \mathbf{a_N}\}$ ($\{\mathbf{b_1}, \cdots, \mathbf{b_N}\}$) representing the statistical atlases.

$\tilde{}$ When used as a diacritic mark, denotes the relaxed version of the coefficient/atlas.

ω In chapter 3, indexed by k, corresponds to the mixing weight of the Gaussian component k. Indexed by nk relates to the spatially varying mixing weights defined in probabilistic atlases. From chapter 4 onward, when indexed by nl_{j_k}, is defined as $\omega_{nl_{j_k}} = b_{nl} a_{n_j} w_{l_{j_k}}$.

$\mathbf{\Omega}$ In Chapter 3, corresponds to the statistical atlases used for the simple GMM. From Chapter 4 onward, depending on the context, denotes either the set $\{\mathbf{b}, \mathbf{a}, \mathbf{W}\}$ of global mixing weights or the set $\{\mathbf{B}, \mathbf{A}, \mathbf{W}\}$ using the statistical atlases for Level 1 and Level 2.

Atlas adaptation and parameters

$\mathscr{D}$ Denotes the Dirichlet distribution.

$\mathscr{B}$ Denotes the Beta distribution.

ε Defines the strength of the prior relaxation process. A lower value represents a stronger relaxation. From chapter 4, in the three-layered hierarchical framework, is indexed by the level of the hierarchy (1 or 2).

κ Directly related to ε. The highest κ, the strongest the relaxation.

ϖ_{nk}, $\boldsymbol{\varpi_n}$ In chapter 3 denote the Dirichlet prior parameters at voxel n for the component k of the GMM and the corresponding vector of gathered values

$\alpha_{nj}, \boldsymbol{\alpha}_n$ From chapter 4 onward, similar to $\varpi_{nk}, \boldsymbol{\varpi}_n$ but applied to the J anatomical tissues defined at Level 2 of the three-layered hierarchical model.

$\beta_{nj}, \boldsymbol{\beta}_n$ Same as α_{nj} and $\boldsymbol{\alpha}_n$ but at Level 1 of the hierarchy.

p_{nl} Responsibility marginalized over the second level.

p_{nj} Responsibility marginalized over the first level.

G_σ Gaussian kernel with standard deviation σ.

MRF notations

$\mathbf{N}_n$ Denotes the set of Von Neumann neighbours of voxel n, *i.e.* the 6 nearest neighbours (east, west, north, south, top and bottom).

$\mathbf{p}_n$ Denotes the vector of responsibilities for all l_{j_k} components.

$P_{\mathbf{N}_n}$ Denotes the set of responsibilities vector $\mathbf{p}_m$ where $m \in \mathbf{N}_n$.

U_{MRF} Energy function related to the current labelling configuration.

$\phi_{nl_{j_k}}$ Abbreviation of $f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_{l_{j_k}}, \boldsymbol{\Xi}_{\mathbf{K}}\right)$.

$\psi_{nl_{j_k}}$ Abbreviation of $\exp\left(-U_{\text{MRF}}\left(\mathbf{e}_{l_{j_k}} \middle| \mathbf{P}^{(t)}_{\mathbf{N}_n}, H\right)\right)$.

H MRF inter-class energy matrix.

Constraint over the covariance matrix

$\mathfrak{e}$ Denotes the global noise model on the image.

$\Delta_{l_{j_k}}$ Weighted covariance matrix for the Gaussian component l_{j_k} at Level 3.

Ψ Prior over the model covariances.

$S_{l_{j_k}}$ Scaling diagonal matrix used to compensate for the log-transformation of the intensities in the prior over the covariances.

Model evolution

$\mathfrak{p}_d^{(k)}$ Corresponds to the d^{th} highest eigenvalue of the covariance matrix Λ_k

$\mathbf{q}_d^{(k)}, Q_k$ Vector of the orthogonal decomposition of the covariance matrix Λ_k corresponding to the eigenvalue $\mathfrak{p}_d^{(k)}$ and associated orthogonal matrix.

KLD_S Calculated on k (l_{j_k} in the hierarchical scheme) denotes the Kullback-Leibler divergence for a component to split.

KLD_M Evaluated on the double k_1, k_2, denotes the Kullback-Leibler divergence for the comparison of two components to merge.

$\mathfrak{v}$ Decimation factor accounting for the proportion of truly independent voxels.

$\text{BIC}(\mathbf{K})$ Bayesian Information Criterion on model $\mathbf{K}$.

$\mathscr{P}(\mathbf{K})$ Penalisation function over the model $\mathbf{K}$ used in BIC

corr_r Value of correlation between adjacent voxels in the r-direction.

Lesion definition

Patho Modalities that can be used as indicator of pathology in the case of WMH, that is FLAIR, T2 and PD.

L Set of components potentially considered as lesions.

TL Set of components considered as lesions.

S Set of possible lesion-related components that need further refinement.

RL Final set of lesion-related components refining S.

Correction for false positives

S_T Tissue binary mask after final classification. T takes its values in GM, WM, CSF, NB...

Spatial distribution

S_T Tissue binary mask after final classification. T takes its values in GM, WM, CSF, NB...

Longitudinal extension

h_τ Intensity mapping polynomial coefficients.

$A(\hat{\mathbf{Y}})$ Polynomial matrix of $\hat{\mathbf{Y}}$.

$T_{\tau\to\mathrm{GW}}$ Spatial transformation from time point space τ to groupwise (GW) space.

t Indexed by n it refers to the typicality map.

$M_{\mathbf{h}_\tau}$ Intensity mapping transformation with parameters $\mathbf{h}_\tau$

$\tilde{N}$ Number of degree of freedom of the Inverse Wishart distribution

Longitudinal simulator

L_S Simulated lesion map.

L Lesion map.

I Original image.

S Simulated image.

BF Bias field.

G Gaussianly distributed lesion intensities.

R Rigid transformation.

Bibliography

[1] P. Malloy, S. Correla, G. Stebbins, and D. H. Laldlaw, "Neuroimaging of White Matter in Aging and Dementia," *The Clinical Neuropsychologist*, vol. 21, no. 1, pp. 73–109, 2007.

[2] C. M. Deber and S. J. Reynolds, "Central nervous system myelin: structure, function and pathology," *Clinical Biochemistry*, vol. 24, pp. 113–134, 1991.

[3] J. D. Schmahmann, E. E. Smith, F. S. Eichler, and C. M. Filley, "Cerebral white matter: Neuroanatomy, clinical neurology, and neurobehavioral correlates," *Annals of the New York Academy of Sciences*, vol. 1142, pp. 266–309, 2008.

[4] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. M. B. Breteler, H. Chabriat, C. DeCarli, F.-E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. van Oostenbrugge, L. Pantoni, O. Speck, B. C. M. Stephan, S. Teipel, Viswanathan Anand, D. Werring, C. Chen, C. Smith, M. A. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *The Lancet Neurology*, vol. 12, pp. 822–838, 2013.

[5] B. E. Grueter and U. G. Schulz, "Age-related cerebral white matter disease (leukoaraiosis): a review," *Postgraduate Medical Journal*, vol. 88, pp. 79–87, 2012.

[6] V. C. Hachinski, P. Potter, and H. Merskey, "Leuko-Araiosis," *Archives of Neurology*, vol. 44, pp. 21–23, jan 1987.

[7] R. Hurford, A. Charidimou, Z. Fox, L. Cipolotti, R. Jäger, and D. J. Werring, "MRI-visible perivascular spaces: relationship to cognition and small vessel disease MRI markers in ischaemic stroke and TIA," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 85, pp. 522–525, may 2014.

[8] F. N. Doubal, A. M. J. MacLullich, K. J. Ferguson, M. S. Dennis, and J. M. Wardlaw, "Enlarged perivascular spaces on MRI are a feature of cerebral small vessel disease," *Stroke*, vol. 41, pp. 450–454, mar 2010.

[9] F. M. Gunning-Dixon, A. M. Brickman, J. C. Cheng, and G. S. Alexopoulos, "Aging of cerebral white matter : a review of MRI findings," *International Journal of Geriatric Psychiatry*, 2008.

[10] C. M. Filley, *The behavioral neurology of white matter*. Oxford University Press, second ed., 2012.

[11] Y. Y. Xiong and V. Mok, "Age-Related White Matter Changes," *Journal of Aging Research*, vol. 2011, 2011.

[12] S. B. Wharton, J. E. Simpson, C. Brayne, and P. G. Ince, "Age-Associated White Matter Lesions: The MRC Cognitive Function and Ageing Study," *Brain Pathology*, vol. 25, pp. 35–43, 2015.

[13] S. Black, F. Gao, and J. Bilbao, "Understanding white matter disease: Imaging-pathological correlations in vascular cognitive impairment," *Stroke*, vol. 40, no. 3 SUPPL. 1, pp. 48–53, 2009.

[14] D. Erten-Lyons, R. Woltjer, J. Kaye, N. Mattek, H. H. Dodge, S. Green, H. Tran, D. B. Howieson, K. Wild, L. C. Silbert, R. Woljter, J. Kaye, N. Mattek, H. H. Dodge, S. Green, H. Tran, D. B. Howieson, K. Wild, and L. C. Silbert, "Neuropathologic basis of white matter hyperintensity accumulation with advanced age," *Neurology*, vol. 81, pp. 977–983, 2013.

[15] J. M. Wardlaw, C. Smith, and M. Dichgans, "Mechanisms underlying sporadic cerebral small vessel disease: insights from neuroimaging," *The Lancet Neurology*, vol. 12, may 2013.

[16] R. Schmidt, H. Schmidt, J. Haybaeck, M. Loitfelder, S. Weis, M. Cavalieri, S. Seiler, C. Enzinger, S. Ropele, T. Erkinjuntti, L. Pantoni, P. Scheltens, F. Fazekas, and K. Jellinger, "Heterogeneity in age-related white matter changes," *Acta Neuropathology*, vol. 122, pp. 171–185, jun 2011.

[17] A.-M. Enciu, M. Gherghiceanu, and B. O. Popescu, "Triggers and effectors of oxidative stress at blood-brain barrier level: relevance for brain ageing and neurodegeneration," *Oxidative Medicine and Cellular Longevity*, vol. 2013, 2013.

[18] K. Woong Kim, J. R. MacFall, and M. E. Payne, "Classification of white matter lesions on magnetic resonance imaging in the elderly," *Biological Psychiatry*, vol. 64, pp. 273–290, aug 2008.

[19] A. A. Gouw, A. Seewann, W. M. van der Flier, F. Barkhof, R. A. M., P. Scheltens, and J. J. G. Geurts, "Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, pp. 126–135, 2011.

[20] S. Haller, E. Kövari, F. R. Herrmann, V. Cuvinciuc, A.-M. Tomm, G. B. Zulian, K.-O. Lovblad, P. Giannakopoulos, and C. Bouras, "Do brain T2/FLAIR white matter hyperintensities correspond to myelin loss in normal aging? A radiologic-neuropathologic correlation study," *Acta Neuropathologica Communications*, vol. 1, no. 14, 2013.

[21] M. Yoshita, E. Fletcher, and C. DeCarli, "Current concepts of analysis of cerebral white matter hyperintensities on magnetic resonance imaging.," *Topics in magnetic resonance imaging : TMRI*, vol. 16, pp. 399–407, dec 2005.

[22] P. Maillard, E. Fletcher, S. N. Lockhart, A. E. Roach, B. Reed, D. Mungas, C. DeCarli, and O. T. Carmichael, "White matter hyperintensities and their penumbra lie along a continuum of injury in the aging brain," *Stroke*, vol. 45, pp. 1721–1726, 2014.

[23] P. Maillard, O. Carmichael, D. Harvey, E. Fletcher, B. Reed, D. Mungas, and C. DeCarli, "FLAIR and Diffusion MRI signals are independent predictors of white matter hyperintensities," *American Journal of Neuroradiology*, vol. 34, pp. 54–61, jan 2013.

[24] B. Patel and H. S. Markus, "Magnetic resonance imaging in cerebral small vessel disease and its use as a surrogate disease marker," *International Journal of Stroke*, vol. 6, pp. 47–59, feb 2011.

[25] M. Radanovic, F. Ramos, S. Pereira, F. Stella, I. Aprahamian, L. Kobutl Ferreira, O. V. Forlenza, and G. F. Busatto, "White matter abnormalities associated with Alzheimer's disease and mild cognitive impairment: a critical review of MRI studies," *Expert Reviews Neurotherapeutics*, vol. 13, no. 5, pp. 483–493, 2013.

[26] D. H. Salat, "Imaging small vessel-associated white matter changes in aging," *Neuroscience*, 2014.

[27] E. Rostrup, A. A. Gouw, H. Vrenken, E. C. W. Van Straaten, S. Ropele, L. Pantoni, D. Inzitari, F. Barkhof, and G. Waldemar, "The spatial distribution of age-related white matter changes as a function of vascular risk factors-Results from the LADIS study," *NeuroImage*, vol. 60, no. 3, pp. 1597–1607, 2012.

[28] S. E. Vermeer, M. Hollander, E. J. van Dijk, A. Hofman, P. J. Koudstaal, and M. M. B. Breteler, "Silent brain infarcts and white matter lesions increase stroke risk in the general population: The Rotterdam Scan Study," *Stroke*, vol. 34, pp. 1126–1129, may 2003.

[29] E. Kliper, E. Ben Assayag, R. Tarrasch, M. Artzi, A. D. Korczyn, S. Shenhar-Tsarfaty, O. Aizenstein, H. Hallevi, A. Mike, L. Shopin, N. M. Bornstein, and D. B. Bashat, "Cognitive State following Stroke: The Predominant Role of Pre-existing White Matter Lesions," *PLoS ONE*, vol. 9, no. 8, p. e105461, 2014.

[30] A. C. Birdsill, R. L. Koscik, E. M. Jonaitis, S. C. Johnson, O. C. Okonkwo, B. P. Hermann, A. LaRue, M. A. Sager, B. B. Bendlin, C. Birdsill Alex, R. L. Koscik, E. M. Jonaitis, S. C. Johnson, C. Okonkwo Ozioma, B. P. Hermann, A. LaRue, M. A. Sager, and B. B. Bendlin, "Regional white matter hyperintensities: aging, Alzheimer's disease risk and cognitive function," *Neurobiology of Aging*, vol. 35, no. 4, pp. 769–776, 2014.

[31] M. O'Sullivan, D. K. Jones, P. E. Summers, R. G. Morris, S. C. R. Williams, and H. S. Markus, "Evidence for cortical "disconnection" as a mechanism of age-related cognitive decline," *Neurology*, vol. 57, pp. 632–638, aug 2001.

[32] L. Pantoni and The LADIS Study group, "2001-2011: a decade of the LADIS (Leukoaraiosis And DISability) Study: what have we learned about white matter changes and small-vessel disease," *Cerebrovascular Diseases*, vol. 32, pp. 577–588, 2011.

[33] M. S. Krishnan, J. T. O'Brien, M. J. Firbank, L. Pantoni, G. Carlucci, T. Erkinjuntti, A. Wallin, L.-O. Wahlund, P. Scheltens, E. C. W. Van Straaten, D. Inzitari, and The LADIS Study group, "Relationship between periventricular and deep white matter lesions and depressive symptoms in older people. The LADIS study," *International Journal of Geriatric Psychiatry*, vol. 21, pp. 1470–1477, oct 2006.

[34] M. Mortamais, C. Reynes, A. M. Brickman, F. A. Provenzano, J. Muraskin, F. Portet, C. Berr, J. Touchon, A. Bonafé, E. le Bars, J. J. Maller, C. Meslin, R. Sabatier, K. Ritchie, and S. Artero, "Spatial distribution of cerebral white matter lesions predicts progression to mild cognitive impairment and dementia," *PLoS ONE*, vol. 8, no. 2, 2013.

[35] N. Bolandzadeh, J. C. Davis, R. Tam, T. C. Handy, and T. Liu-Ambrose, "The association between cognitive function and white matter lesion location in older adults: a systematic review," *BMC Neurology*, vol. 12, no. 126, 2012.

[36] C. DeCarli, E. Fletcher, V. Ramey, D. Harvey, and W. J. Jagust, "Anatomical mapping of white matter hyperintensities (WMH): exploring the relationships between periventricular WMH, deep WMH and total WMH burden," *Stroke*, vol. 36, pp. 50–55, 2005.

[37] M. Mortamais, S. Artero, and K. Ritchie, "Cerebral white matter hyperintensities in the prediction of cognitive decline and incident dementia," *International Review of Psychiatry*, vol. 25, pp. 686–698, dec 2013.

[38] C. A. Raji, O. L. Lopez, L. H. Kuller, O. T. Carmichael, W. T. J. Longstreth, M. H. Gach, J. Boardmann, C. B. Bernick, P. M. Thomson, and J. T. Becker, "White matter lesions and brain gray matter volume in cognitively normal elders," *Neurobiology of Aging*, vol. 33, 2012.

[39] M. Haglund and E. Englund, "Cerebral amyloid angiopathy, white matter lesions and Alzheimer encephalopathy - a histopathological assessment," *Dementia and Geriatric Cognitive Disorders*, vol. 14, pp. 161–166, jan 2002.

[40] R. Schmidt, S. Ropele, C. Enzinger, K. Petrovic, S. Smith, H. Schmidt, P. M. Matthews, and F. Fazekas, "White matter lesion progression, brain atrophy, and cognitive decline: the Austrian stroke prevention study," *Annals of Neurology*, vol. 58, pp. 610–6, oct 2005.

[41] M. C. Power, J. A. Deal, A. R. Sharrett, C. R. Jack, D. Knopman, T. H. Mosley, and R. F. Gottesman, "Smoking and white matter hyperintensity progression The ARIC-MRI Study," *Neurology*, vol. 84, pp. 841–848, 2015.

[42] W. D. Taylor, D. C. Steffens, J. R. MacFall, D. R. McQuoid, M. E. Payne, J. M. Provenzale, and K. R. R. Krishnan, "White matter hyperintensity progression and late-life depression outcomes," *Archives of General Psychiatry*, vol. 60, pp. 1090–1096, nov 2003.

[43] F. Fazekas, H. Offenbacher, S. Fuchs, R. Schmidt, K. Niederkorn, S. Horner, and H. Lechner, "Criteria for an increased specificity of MRI interpretation in elderly subjects with suspected multiple sclerosis," *Neurology*, vol. 38, pp. 1822–1825, dec 1988.

[44] M. E. Murray, P. Vemuri, G. M. Preboske, M. C. Murphy, K. J. Schweitzer, J. E. Parisi, C. R. Jack, and D. W. Dickson, "A quantitative postmortem MRI design sensitive to white matter hyperintensity differences and their relationship with underlying pathology.," *Journal of Neuropathology and Experimental Neurology*, vol. 71, pp. 1113–1122, dec 2012.

[45] E. C. Leritz, J. Shepel, V. J. Williams, L. A. Lipsitz, R. E. McGlinchey, W. P. Milberg, and D. H. Salat, "Associations between T1 white matter lesion volume and regional white matter microstructure in aging," *Human Brain Mapping*, vol. 35, no. 3, pp. 1085–1100, 2014.

[46] E. Olsson, N. Klasson, J. Berge, C. Eckerström, A. Edman, H. Malmgren, and A. Wallin, "White matter lesion assessment in patients with cognitive impairment and healthy controls: reliability comparisons between visual rating, a manual and an automatic volumetrical MRI method- The Gothenburg MCI study," *Journal of Aging Research*, vol. 2013, 2013.

[47] J. C. de Groot, F. E. de Leeuw, M. Oudkerk, J. van Gijn, A. Hofman, J. Jolles, and M. M. Breteler, "Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study.," *Annals of Neurology*, vol. 47, pp. 145–51, feb 2000.

[48] M. Rovaris, G. Comi, M. A. Rocca, M. Cercignani, B. Colombo, G. Santuccio, and M. Filippi, "Relevance of hypointense lesions on fast fluid attenuated inversion recovery MR images as marker of disease severity in cases of multiple sclerosis," *American Journal of Neuroradiology*, vol. 20, pp. 813–820, 1999.

[49] F. Admiraal-Behloul, D. M. J. M. van den Heuvel, H. Olofsen, M. J. P. van Osch, J. van der Grond, M. a. Van Buchem, and J. H. C. Reiber, "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly," *NeuroImage*, vol. 28, pp. 607–617, 2005.

[50] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, À. Rovira, L. Ramió-Torrent, and L. Rovira, "Segmentation of multiple sclerosis lesions in brain MRI : a review of automated approaches," *Information Sciences*, vol. 186, pp. 164–185, 2012.

[51] D. Uhlenbrock and S. Sehlen, "The value of T1-weighted images in the differentiation between MS, white matter lesions, and subcortical arteriosclerotic encephalopathy (SAE)," *Neuroradiology*, vol. 31, pp. 203–212, 1989.

[52] W. Zhan, Y. Zhang, S. G. Mueller, P. Lorenzen, S. Hadjidemetriou, N. Schuff, and M. W. Weiner, "Characterization of white matter degeneration in elderly subjects by magnetic resonance diffusion and FLAIR imaging correlation," *NeuroImage*, vol. 48, pp. 758–765, 2009.

[53] E. Lavdas, I. Tsougos, S. Kogia, G. Gratsias, P. Svolos, V. Roka, I. V. Fezoulidis, and E. Kapsalaki, "T2 FLAIR artifacts at 3-T brain magnetic resonance imaging," *Clinical Imaging*, vol. 38, pp. 85–90, jan 2014.

[54] R. Bakshi, S. D. Caruthers, V. Janardhan, and M. Wasay, "Intraventricular CSF Pulsation Artifact on Fast Fluid-Attenuated Inversion-Recovery MR Images: Analysis of 100 Consecutive Normal Studies," *American Journal of Neuroradiology*, vol. 21, pp. 503–508, mar 2000.

[55] S. Kakeda, Y. Korogi, Y. Hiai, N. Ohnari, T. Sato, and T. Hirai, "Pitfalls of 3D FLAIR Brain Imaging," *Academic Radiology*, vol. 19, pp. 1225–1232, 2012.

[56] M. Neema, Z. D. Guss, J. M. Stankiewicz, A. Arora, B. C. Healy, and R. Bakshi, "Normal findings on brain fluid attenuated inversion recovery MR images at 3T," *American Journal of Neuroradiology*, vol. 30, pp. 911–916, may 2009.

[57] M. I. Gawne-Cain, N. C. Silver, I. F. Moseley, and D. H. Miller, "Fast FLAIR of the brain: the range of appearances in normal subjects and its application to quantification of white-matter disease," *Neuroradiology*, vol. 39, pp. 243–249, 1997.

[58] M. D. C. Valdés Hernández, Z. Morris, D. A. Dickie, N. a. Royle, S. Muñoz Maniega, B. S. Aribisala, M. E. Bastin, I. J. Deary, J. M. Wardlaw, M. d. Valdès Hernández, Z. Morris, D. A. Dickie, N. a. Royle, S. Muñoz Maniega, B. S. Aribisala, M. E. Bastin, I. J. Deary, and J. M. Wardlaw, "Close correlation between quantitative and qualitative assessments of white matter lesions," *Neuroepidemiology*, vol. 40, pp. 13–22, 2013.

[59] M. E. Payne, D. L. Fetzer, J. R. MacFall, J. M. Provenzale, C. E. Byrum, and K. R. R. Krishnan, "Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects," *Psychiatry Research: Neuroimaging*, vol. 115, pp. 63–77, 2002.

[60] P. Maillard, E. Fletcher, D. Harvey, O. Carmichael, B. Reed, D. Mungas, and C. DeCarli, "White matter hyperintensity penumbra," *Stroke*, vol. 42, pp. 1917–1922, jun 2011.

[61] M. Valdès Hernández, R. J. Piper, M. E. Bastin, N. A. Royle, B. S. Maniega, C. Murray, I. J. Deary, and J. M. Wardlaw, "Morphologic, Distributional, Volumetric and Intensity characterization of periventricular hyperintensities," *American Journal of Neuroradiology*, pp. 1–8, 2013.

[62] E. C. W. Van Straaten, F. Fazekas, E. Rostrup, P. Scheltens, R. Schmidt, L. Pantoni, D. Inzitari, G. Waldemar, T. Erkinjuntti, R. Mäntylä, L.-O. O. Wahlund, and F. Barkhof, "Impact of white matter hyperintensities scoring method on correlations with clinical data: the LADIS study.," *Stroke*, vol. 37, pp. 836–840, mar 2006.

[63] J. M. Wardlaw, K. J. Ferguson, and C. Graham, "White matter hyperintensities and rating scales - Observer reliability varies with lesion load," *Journal of Neurology*, vol. 251, no. 5, pp. 584–590, 2004.

[64] R. Mäntylä, T. Erkinjuntti, O. Salonen, H. J. Aronen, T. Peltonen, T. Pohjasvaara, and C.-G. Standertskjöld-Nordenstam, "Variable agreement between visual rating scales for white matter hyperintensities on MRI," *Stroke*, vol. 28, no. 8, pp. 1614–1623, 1997.

[65] L. Pantoni, M. Simoni, G. Pracucci, R. Schmidt, F. Barkhof, and D. Inzitari, "Visual rating scales for age-related white matter changes (leukoaraiosis): Can the heterogeneity be reduced?," *Stroke*, vol. 33, no. 12, pp. 2827–2833, 2002.

[66] P. Kapeller, R. Barber, R. J. Vermeulen, H. Adèr, P. Scheltens, W. Freidl, O. Almkvist, M. Moretti, T. del Ser, P. Vaghfeldt, C. Enzinger, F. Barkhof, D. Inzitari, T. Erkinjunti, R. Schmidt, F. Fazekas, H. Ader, P. Scheltens, W. Freidl, O. Almkvist, M. Moretti, T. del Ser, P. Vaghfeldt, C. Enzinger, F. Barkhof, D. Inzitari, T. Erkinjuntti, R. Schmidt, and F. Fazekas, "Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement and correlations with quantitative measurements," *Stroke*, vol. 34, no. 2, pp. 441–445, 2003.

[67] D. M. J. van den Heuvel, V. H. ten Dam, A. J. M. de Craen, F. Admiraal-Behloul, A. van Es, W. M. Palm, A. Spilt, E. Bollen, G. J. Blauw, L. J. Launer, R. G. J. Westendrop, and M. A. van Buchem, "Measuring longitudinal white matter changes: comparison of a visual rating scale with a volumetric measurement," *American Journal of Neuroradiology*, vol. 27, pp. 875–878, 2006.

[68] M. Battaglini, M. Jenkinson, and N. De Stefano, "Evaluating and Reducing the Impact of White Matter Lesions on Brain Volume Measurements," *Human Brain Mapping*, vol. 33, pp. 2062–2071, 2012.

[69] R. Gelineau-Morel, V. Tomassini, M. Jenkinson, H. Johansen-Berg, P. M. Matthews, and J. Palace, "The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis," *Human Brain Mapping*, vol. 33, pp. 2802–2814, 2012.

[70] N. Levy-Cooperman, J. Ramirez, N. J. Lobaugh, and S. E. Black, "Misclassified tissue volumes in Alzheimer Disease Patients with white matter hyperintensities: importance of lesion segmentation procedures for volumetric analysis," *Stroke*, vol. 39, pp. 1134–1141, 2008.

[71] E. A. Ashton, C. Takahashi, M. J. Berg, A. Goodman, S. Totterman, and S. Ekholm, "Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI," *Journal of Magnetic Resonance Imaging*, vol. 17, pp. 300–308, mar 2003.

[72] M. Filippi, M. A. Horsfield, S. Bressi, V. Martinelli, C. Baratti, P. Reganati, A. Campi, D. H. Miller, and G. Comi, "Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis," *Brain*, vol. 118, pp. 1593–1600, dec 1995.

[73] D. A. Wicks, P. S. Tofts, D. H. Miller, G. H. du Boulay, A. Feinstein, R. P. Sacares, I. Harvey, R. Brenner, and W. I. McDonald, "Volume measurement of multiple sclerosis lesions with magnetic resonance images. A preliminary study.," *Neuroradiology*, vol. 34, pp. 475–479, jan 1992.

[74] M. Iorio, G. Spalletta, C. Chiapponi, G. Luccichenti, C. Cacciari, M. D. Orfei, C. Caltagirone, and F. Piras, "White matter hyperintensities segmentation: a new semi-automated method.," *Frontiers in Aging Neuroscience*, vol. 5, p. 76, jan 2013.

[75] F. Derraz, L. Peyrodie, A. Pinti, A. Taleb-Ahmed, A. Chikh, and P. Hautecoeur, "Semi-automatic segmentation of Multiple Sclerosis Lesion based active contours model and variational dirichlet process," *CMES - Computer Modeling in Engineering and Sciences*, vol. 67, no. 2, pp. 95–116, 2010.

[76] M. Ghazel, A. Traboulsee, and R. Ward, "Semi-Automated Segmentation of Multiple Sclerosis Lesions in Brain MRI using Texture Analysis," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 6–10, IEEE, aug 2006.

[77] T. Heinonen, P. Dastidar, H. Eskola, H. Frey, P. Ryymin, and E. Laasonen, "Applicability of semi-automatic segmentation for volumetric analysis of brain lesions," *Journal of Medical Engineering & Technology*, vol. 22, pp. 173–178, jul 2009.

[78] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," *NeuroImage*, vol. 49, no. 2, pp. 1524–1535, 2010.

[79] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 677–688, aug 2001.

[80] P. Anbeek, K. L. Vincken, M. J. P. van Osch, R. H. C. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, pp. 1037–1044, 2004.

[81] T. B. Dyrby, E. Rostrup, W. F. Baaré, E. C. van Straaten, F. Barkhof, H. Vrenken, S. Ropele, R. Schmidt, T. Erkinjuntti, L.-O. Wahlund, L. Pantoni, D. Inzitari, O. B. Paulson, L. K. Hansen, and G. Waldemar, "Segmentation of age-related white matter changes in a clinical multi-center study," *NeuroImage*, vol. 41, pp. 335–345, jun 2008.

[82] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, and C. Davatzikos, "Computer Assisted segmentation of white matter lesions in 3D MR images, using support vector machines," *Academic Radiology*, vol. 15, no. 3, pp. 300–313, 2008.

[83] D. Yamamoto, H. Arimura, S. Kakeda, T. Magome, Y. Yamashita, F. Toyofuku, M. Ohki, Y. Higashida, and Y. Korogi, "Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine," *Computerized Medical Imaging and Graphics*, vol. 34, no. 5, pp. 404–413, 2010.

[84] H. Khastavaneh and H. Haron, "False Positives Reduction on Segmented Multiple Sclerosis Lesions Using Fuzzy Inference System by Incorporating Atlas Prior Anatomical Knowledge : A Conceptual Model," *Computational Collective Intelligence. Technologies and Applications*, pp. 11–19, 2014.

[85] M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. C. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "BOOST: a supervised approach for multiple sclerosis lesion segmentation," *Journal of Neuroscience Methods*, vol. 237, pp. 108–17, nov 2014.

[86] O. Freifeld, H. Greespan, and J. Goldberger, "Multiple Sclerosis Lesion Detection Using Constrained GMM and Curve Evolution," *International Journal of Biomedical Imaging*, vol. 2009, p. 13, 2009.

[87] X. Tomas-Fernandez and S. K. Warfield, "A Model of Population and Subject (MOPS) Intensities with Application to Multiple Sclerosis Lesion Segmentation," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1349–1361, jan 2015.

[88] D. García-Lorenzo, J. Lecoeur, D. L. Arnold, D. L. Collins, and C. Barillot, "Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5762 LNCS, pp. 584–591, 2009.

[89] R. Simões, C. Mönninghoff, M. Dlugaj, C. Weimar, I. Wanke, A.-M. M. van Cappellen van Walsum, C. Slump, R. Simoes, Monninghoff Christoph, M. Dlugaj, C. Weimar, I. Wanke, A.-M. M. van Cappellen van Walsum, and C. Slump, "Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images," *Magnetic resonance imaging*, vol. 31, pp. 1182–1189, sep 2013.

[90] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau, "An automated tool for detection of FLAIR-hyperintense white matter lesions in multiple sclerosis," *NeuroImage*, vol. 59, pp. 3774–3783, 2012.

[91] E. Geremia, B. H. Menze, O. Clatz, E. Komukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel MR images," in *MICCAI 2010* (T. Jiang, ed.), LNCS 6361, pp. 111–118, Springer-Verlag, 2010.

[92] Y. Wu, S. K. Warfield, I. L. Tan, W. M. Wells III, D. S. Meier, R. A. van Schijndel, F. Barkhof, and C. R. G. Guttmann, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI," *NeuroImage*, vol. 32, pp. 1205–1215, 2006.

[93] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans, "Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images," *IEEE Transactions on Medical Imaging*, vol. 14, pp. 442–53, jan 1995.

[94] K. H. Ong, D. Ramachandram, R. Mandava, and I. L. Shuaib, "Automatic white matter lesion segmentation using an adaptive outlier detection method," *Magnetic resonance imaging*, vol. 30, no. 6, pp. 807–823, 2012.

[95] T. Samaille, L. Fillon, R. Cuingnet, E. Jouvent, H. Chabriat, D. Dormont, O. Colliot, and M. Chupin, "Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation," *PLoS ONE*, vol. 7, no. 11, 2012.

[96] S. Bricq, C. Collet, and J.-P. Armspach, "Markovian segmentation of 3D brain MRI to detect Multiple Sclerosis lesions," in *2008 15th IEEE International Conference on Image Processing*, pp. 733–736, IEEE, 2008.

[97] L. S. Aït-Ali, S. Prima, P. Hellier, B. Carsin, G. Edan, and C. Barillot, "STREM: A robust multidimensional parametric method to segment MS lesions in MRI," in *MICCAI 2005*, LNCS 3749, pp. 409–416, Springer International, 2005.

[98] X. Wei, S. K. Warfield, K. H. Zou, Y. Wu, X. Li, A. Guimond, J. P. Mugler, R. R. Benson, L. Wolfson, H. L. Weiner, and C. R. G. Guttmann, "Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy," *Journal of Magnetic Resonance Imaging*, vol. 15, pp. 203–209, feb 2002.

[99] S. Datta and N. P. A., "A comprehensive approach to the segmentation of multichannel three-dimensional MR brain images in multiple sclerosis," *NeuroImage: Clinical*, vol. 2, pp. 184–196, 2013.

[100] F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, and M. Dojat, "Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation," in *2010 IEEE International Symposium onBiomedical Imaging: From Nano to Macr*, pp. 69–72, IEEE, 2010.

[101] D. García-Lorenzo, S. Prima, A. L. Douglas, D. L. Collins, and C. Barillot, "Trimmed-Likelihood Estimation for Focal Lesions and Tissue Segmentation in Multisequence MRI for Multiple Sclerosis," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 1455–1467, aug 2011.

[102] H. Vrenken, M. Jenkinson, M. A. Horsfield, M. Battaglini, R. A. van Schijndel, E. Rostrup, J. J. G. Geurts, E. Fisher, A. Zijdenbos, J. Ashburner, D. H. Miller, M. Filippi, F. Fazekas, M. Rovaris, À. Rovira, F. Barkhof, N. De Stefano, and M. S. Group, "Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis," *Journal of Neurology*, vol. 260, pp. 2458–2471, 2013.

[103] C. R. Jack Jr, P. C. O'Brien, D. W. Rettman, M. M. Shiung, X. Yuecheng, R. Muthupillai, A. Manduca, R. Avula, B. J. Erickson, C. R. Jack, P. C. O'Brien, D. W. Rettman, M. M. Shiung, Y. Xu, R. Muthupillai, A. Manduca, R. Avula, and B. J. Erickson, "FLAIR Histogram segmentation for measurement of leukoaraiosis volume," *Journal of Magnetic Resonance Imaging*, vol. 14, pp. 668–676, 2001.

[104] Y. Zhong, D. Utriainen, Y. Wang, Y. Kang, and E. Haacke, "Automated white matter hyperintensity detection in multiple sclerosis using 3D T2 FLAIR," *International Journal of Biomedical Imaging*, 2014.

[105] B. I. Yoo, J. J. Lee, J. W. Han, S. Y. W. Oh, E. Y. Lee, J. R. MacFall, M. E. Payne, T. H. Kim, J. H. Kim, and K. W. Kim, "Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images," *Neuroradiology*, vol. 56, pp. 265–281, 2014.

[106] A. Khademi, A. Venetsanopoulos, and A. R. Moody, "Robust white matter lesion segmentation in FLAIR MRI," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 860–871, mar 2012.

[107] R. Khayati, M. Vafadust, F. Towdhidkhah, and S. M. Nabavi, "Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model," *Computers in Biology and Medicine*, vol. 38, pp. 379–390, 2008.

[108] J. A. Maldjian, C. T. Whitlow, B. N. Saha, G. Kota, C. Vandergriff, E. M. Davenport, J. Divers, B. I. Freedman, and D. W. Bowden, "Automated white matter total lesion volume segmentation in diabetes," *American Journal of Neuroradiology*, vol. 34, pp. 2265–2270, 2013.

[109] D. Mortazavi, A. Z. Kouzani, H. Soltanian-Zadeh, and H. Soltanian-Zaheh, "Segmentation of multiple sclerosis lesions in MR images : a review," *Diagnostic Neuroradiology*, vol. 54, pp. 299–320, 2012.

[110] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, vol. 17, pp. 1–18, 2013.

[111] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging : A Review," *Neuroinformatics*, vol. 13, pp. 261–276, jul 2015.

[112] N. Guizard, P. Coupé, V. S. Fonov, J. V. Manjón, D. L. Arnold, and D. L. Collins, "Rotation-invariant multi-contrast non-local means for MS lesion segmentation.," *NeuroImage: Clinical*, vol. 8, pp. 376–389, jan 2015.

[113] R. de Boer, H. A. Vrooman, M. A. Ikram, M. W. Vernooij, M. M. B. Breteler, A. van der Lugt, and W. J. Niessen, "Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods," *NeuroImage*, vol. 51, pp. 1047–1056, jul 2010.

[114] L. J. Patino-Correa, O. Pogrebnyak, J. A. Martinez-Castro, and E. M. Felipe-Riveron, "White matter hyper-intensities automatic identification and segmentation in magnetic resonance imaging," *Expert systems with applications*, vol. 41, pp. 7114–7123, nov 2014.

[115] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson, "Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies," *Human Brain Mapping*, vol. 35, pp. 4219–4235, aug 2014.

[116] R. Mechrez, J. Goldberger, and H. Greenspan, "MS lesion segmentation using a multi-channel patch-based approach with spatial consistency," in *SPIE Medical Imaging* (S. Ourselin and M. A. Styner, eds.), p. 94130O, International Society for Optics and Photonics, mar 2015.

[117] B. R. Sajja, S. Datta, R. He, M. Mehta, R. K. Gupta, J. S. Wolinsky, N. P. A., and P. A. Narayana, "Unified approach for Multiple Sclerosis Lesion Segmentation on Brain MRI," *Annals of Biomedical Engineering*, vol. 34, pp. 142–151, jan 2006.

[118] M. D. M. D. Steenwijk, P. J. W. Pouwels, M. Daams, J. W. van Dalen, M. W. A. Caan, E. Richard, F. Barkhof, and H. Vrenken, "Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)," *NeuroImage: Clinical*, vol. 3, pp. 462–469, 2013.

[119] T. Samaille, O. Colliot, D. Dormont, and M. Chupin, "Automatic segmentation of age-related white matter changes on FLAIR images: methods and multicentre validation," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, IEEE, 2011.

[120] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó, "A toolbox for multiple sclerosis lesion segmentation.," *Neuroradiology*, vol. 57, pp. 1031–1043, oct 2015.

[121] Y. Wango, J. A. Catindig, S. Hilal, H. W. Soon, E. Ting, T. Y. Wong, N. Venketasubramanian, C. Chen, A. Qiu, Y. Wang, J. A. Catindig, S. Hilal, H. W. Soon, E. Ting, T. Y. Wong, N. Venketasubramanian, C. Chen, and A. Qiu, "Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts," *NeuroImage*, vol. 60, no. 4, pp. 2379–2388, 2012.

[122] A. P. Zijdenbos, R. Forghani, and A. C. Evans, "Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 1280–1291, oct 2002.

[123] R. Wang, C. Li, J. Wang, X. Wei, Y. Li, C. Hui, Y. Zhu, and S. Zhang, "Automatic segmentation of white matter lesions on magnetic resonance images of the brain by using an outlier detection strategy," *Magnetic Resonance Imaging*, vol. 32, pp. 1321–1329, dec 2014.

[124] R. Wang, C. Li, J. Wang, X. Wei, Y. Li, C. Hui, Y. Zhu, and S. Zhang, "Automatic Segmentation and Quantitative Analysis of White Matter Hyperintensities on FLAIR Images Using Trimmed-Likelihood Estimator," *Academic Radiology*, vol. 21, pp. 1512–1523, dec 2014.

[125] B. Gaonkar, G. Erus, N. Bryan, and C. Davatzikos, "Automated segmentation of brain lesions by combining intensity and spatial information," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 93–96, IEEE, 2010.

[126] J.-C. Souplet, C. Lebrun, N. Ayache, and G. Malandain, "An automatic segmentation of T2-FLAIR Multiple sclerosis lesions," in *MICCAI-Multiple sclerosis lesion segmentation challenge workshop*, 2008.

[127] G. Dugas-Phocion, M. A. Gonzalez Ballester, C. Lebrun, S. Chanalet, C. Bensa, G. Malandain, and N. Ayache, "Hierarchical Segmentation of Multiple Sclerosis Lesions in Multi-Sequence MRI," in *International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'04)*, apr 2004.

[128] R. Schmidt, C. Enzinger, S. Ropele, H. Schmidt, and F. Fazekas, "Progression of cerebral white matter lesions: 6-year results of the Austrian Stroke Prevention Study," *The Lancet*, vol. 361, pp. 2046–2048, jun 2003.

[129] A. Carass, J. Cuzzocreo, M. B. Wheeler, P.-L. Bazin, S. M. Resnick, and J. L. Prince, "Simple paradigm for extra-cerebral tissue removal: algorithm and analysis," *NeuroImage*, vol. 56, pp. 1982–1992, jun 2011.

[130] R. Stokking, K. L. Vincken, and M. A. Viergever, "Automatic morphology-based brain segmentation (MBRASE) from MRI-T1 data," *NeuroImage*, vol. 12, pp. 726–738, dec 2000.

[131] Z. Hou, "A review on MR Image Intensity Inhomogeneity Correction," *International Journal of Biomedical Imaging*, vol. 2006, pp. 1–11, feb 2006.

[132] B. Johnston, M. S. Atkins, B. Mackiewich, and M. Anderson, "Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 154–169, apr 1996.

[133] J. Gao, C. Li, C. Feng, M. Xie, Y. Yin, and C. Davatzikos, "Non-locally regularized segmentation of multiple sclerosis lesion from multi-channel MRI data," *Magnetic Resonance Imaging*, vol. 32, pp. 1058–1066, oct 2014.

[134] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[135] A.-O. O. Boudraa, S. M. Réda Dehak, Y.-M. M. Zhu, C. Pachai, Y.-G. G. Bao, J. Grimaud, S. M. Dehak, Y.-M. M. Zhu, C. Pachai, Y.-G. G. Bao, and J. Grimaud, "Automated segmentation of multiple sclerosis lesions in multispectral MR imaging using fuzzy clustering," *Computers in Biology and Medicine*, vol. 30, pp. 23–40, 2000.

[136] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[137] A. Evans, D. Collins, S. Mills, E. Brown, R. Kelly, and T. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, pp. 1813–1817, 1993.

[138] A. Jog, A. Carass, D. L. Pham, and J. L. Prince, "Multi-Output Decision Trees for Lesion Segmentation in Multiple Sclerosis a Image," in *SPIE Medical Imaging* (S. Ourselin and M. A. Styner, eds.), vol. 9413, pp. 1–6, International Society for Optics and Photonics, mar 2015.

[139] E. M. Sweeney, R. T. Shinohara, N. Shiee, F. J. Mateen, A. A. Chudgar, J. L. Cuzzocreo, P. A. Calabresi, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, "OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI," *NeuroImage: Clinical*, vol. 2, pp. 402–413, 2013.

[140] S. Damangir, A. Manzouri, K. Oppedal, S. Carlsson, M. J. Firbank, H. Sonnesyn, O. B. Tysnes, J. T. O'Brien, M. K. Beyer, E. Westman, D. Aarsland, L. O. Wahlund, and G. Spulber, "Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines," *Journal of the Neurological Sciences*, vol. 322, no. 1-2, pp. 211–216, 2012.

[141] M. Sdika and D. Pelletier, "Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping," *Human Brain Mapping*, vol. 30, pp. 1060–1067, apr 2009.

[142] E. M. Sweeney, J. T. Vogelstein, J. L. Cuzzocreo, P. A. Calabresi, D. S. Reich, C. M. Crainiceanu, and R. T. Shinohara, "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI.," *PloS One*, vol. 9, p. e95753, jan 2014.

[143] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: method and validation," *IEEE Transactions on Medical Imaging*, vol. 13, pp. 716–724, jan 1994.

[144] H. Deshpande, P. Maurel, and C. Barillot, "Detection of Multiple Sclerosis Lesions using Sparse Representations and Dictionary Learning," in *2nd International Workshop on Sparsity Techniques in Medical Imaging (STMI), MICCAI 2014*, no. 71-79, pp. 1–9, sep 2014.

[145] S. Roy, Q. He, A. Carass, A. Jog, J. L. Cuzzocreo, D. S. Reich, J. Prince, and D. Pham, "Example based lesion segmentation," in *SPIE Medical Imaging* (S. Ourselin and M. A. Styner, eds.), p. 90341Y, International Society for Optics and Photonics, mar 2014.

[146] M. Anitha, P. T. Selvy, and V. Palanisamy, "automated detection of white matter lesions in MRI brain images using spatio-fuzzy and spatio-possibilistic clustering models," *Computer Science & Engineering*, vol. 2, no. 2, pp. 1–11, 2012.

[147] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[148] P. Schroeter, J.-M. M. Vesin, T. Langenberger, and R. Meuli, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 172–186, apr 1998.

[149] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm.," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[150] S. Warfield, J. Dengler, J. Zaers, C. R. Guttmann, W. M. Wells, G. J. Ettinger, J. Hiller, and R. Kikinis, "Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions," *Journal of Image Guided Surgery*, vol. 1, no. 6, pp. 326–338, 1995.

[151] R. Harmouche, D. L. Collins, D. Arnold, S. Francis, and T. Arbel, "Bayesian MS Lesion Classification Modeling Regional and Local Spatial Information," in *18th International Conference on Pattern Recognition (ICPR'06)*, 2006.

[152] A. Galimzianova, F. Pernuš, B. Likar, and Ž. Špiclin, "Stratified mixture modeling for segmentation of white-matter lesions in brain MR images," *NeuroImage*, vol. 124, pp. 1031–1043, jan 2016.

[153] M. Cabezas, A. Oliver, E. Roura, J. Freixenet, J. C. Vilanova, L. Ramió-Torrentà, A. Rovira, and X. Lladó, "Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding," *Computer Methods and Programs in Biomedicine*, vol. 115, pp. 147–161, jul 2014.

[154] G. Dugas-Phocion, M. A. G. Ballester, G. Malandain, C. Lebrun, and N. Ayache, "Improved EM-Based Tissue Segmentation and Partial Volume Effect Quantification in Multi-sequence Brain MRI," in *Medical Image Computing and Computer-Assisted Intervention*, pp. 26–33, 2004.

[155] D. García-Lorenzo, S. Prima, D. L. Collins, D. Arnold, S. P. Morrissey, and C. Barillot, "Combining robust expectation maximization and mean shift algorithms for multiple sclerosis brain segmentation," in *MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)(MIAMS'2008)*, pp. 82–91, 2008.

[156] E. Gibson, F. Gao, S. E. Black, and N. J. Lobaugh, "Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T," *Journal of Magnetic Resonance Imaging*, vol. 31, pp. 1311–1322, jun 2010.

[157] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra, "A multidimensional segmentation evaluation of medical image data," *Computer Methods and Programs in Biomedicine*, vol. 96, pp. 108–124, 2009.

[158] P. Maillard, N. Delcroix, F. Crivello, C. Dufouil, S. Gicquel, M. Joliot, N. Tzourio-Mazoyer, A. Alpérovitch, C. Tzourio, and B. Mazoyer, "An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases," *Neuroradiology*, vol. 50, pp. 31–42, jan 2008.

[159] M. A. Styner, J. Lee, B. Chin, M. S. Chin, O. Commowick, H.-H. Tran, S. Markovic-Plese, V. Jewells, and S. K. Warfield, "3D segmentation in the clinic: a grand challenge II: MS lesion segmentation," in *The MIDAS Journal - MS Lesion Segmentation*, 2008.

[160] N. Lummel, V. Schoepf, M. Burke, H. Brueckmann, and J. Linn, "3D fluid-attenuated inversion recovery imaging: Reduced CSF artifacts and enhanced sensitivity and specificity for subarachnoid hemorrhage," *American Journal of Neuroradiology*, vol. 32, no. 11, pp. 2054–2060, 2011.

[161] S. Naganawa, T. Koshikawa, T. Nakamura, H. Kawai, H. Fukatsu, T. Ishigaki, T. Komada, K. Maruyama, and O. Takizawa, "Comparison of flow artifacts between 2D-FLAIR and 3D-FLAIR sequences at 3 T," *European Radiology*, vol. 14, pp. 1901–1908, oct 2004.

[162] G. McLachlan, T. Krishnan, and S. K. Ng, "The EM algorithm." Papers/ Humboldt-Universitat Berlin, Center for Applied Statistics Economics, 2004.

[163] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *Medical Imaging, IEEE Transactions on*, vol. 18, pp. 885–896, oct 1999.

[164] S. Prima, N. Ayache, T. Barrick, and N. Roberts, "Maximum Likelihood estimation of the bias field in MR brain images: investigating different modelings of the imaging process," in *MICCAI 2001*, pp. 811–819, 2001.

[165] C. Fraley and E. Raftery Adrian, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, jun 2002.

[166] M. Cabezas, A. Oliver, X. Lladó, and J. Freixenet, "A review of atlas-based segmentation for magnetic resonance brain images," *Computer Methods and Programs in Biomedicine*, vol. 104, pp. 158–177, dec 2011.

[167] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, pp. 839–851, jul 2005.

[168] M. J. Cardoso, A. Melbourne, G. S. Kendall, M. Modat, N. J. Robertson, N. Marlow, and S. Ourselin, "AdaPT: an adaptive preterm segmentation algorithm for neonatal brain MRI," *NeuroImage*, vol. 65, pp. 97–108, 2013.

[169] N. Shiee, P.-L. Bazin, J. L. Cuzzocreo, A. Blitz, and D. L. Pham, "Segmentation of brain images using adaptive atlases with application to ventriculomegaly," in *Biennial International Conference on Information Processing in Medical Imaging*, pp. 1–12, Springer, 2011.

[170] J. Zhang, "The Mean Field Theory in EM procedures for Markov Random Fields," *IEEE Transactions on Signal Processing*, vol. 40, pp. 2570–2583, oct 1992.

[171] S. Roy, A. Carass, P.-L. Bazin, and J. L. Prince, "A Rician mixture model classification algorithm for magnetic resonance images," in *Biomedical Imaging: From Nano to Macro*, ISBI'09 IEEE International Symposium on, (Piscataway, NJ, USA), pp. 406–409, IEEE Press, 2009.

[172] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "SMEM Algorithm for Mixture Models," *Neural Computation*, vol. 12, pp. 2109–2128, sep 2000.

[173] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Networks*, vol. 11, pp. 271–282, 1998.

[174] K. Blekas and I. E. Lagaris, "Split-merge incremental learning (SMILE) of mixture models," in *17th International Conference on Artificial Neural Networks*, ICANN'07, Springer, 2007.

[175] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 381–396, mar 2002.

[176] A. Corduneanu and C. M. Bishop, "Variational bayesian model selection for mixture distributions," in *Artificial intelligence and Statistics*, vol. 2001, pp. 27–34, Morgan Kaufmann Waltham, MA, 2001.

[177] B. Zhang, C. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," *Pattern Recognition*, vol. 37, no. 1, pp. 131–144, 2004.

[178] Y. Li and L. Li, "A split and merge EM algorithm for color image segmentation," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, vol. 4, pp. 395–399, nov 2009.

[179] Z. Li, Q. Liu, J. Chen, and H. Lu, "A variational inference based approach for image segmentation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, dec 2008.

[180] J. Li, "Clustering based on a multilayer mixture model," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 547–566, 2004.

[181] G. McLachlan, "On Bootstrapping the Likelihood Ratio Test Statistic for the number of components in a normal mixture," *Journal of the Royal Statistical Society. Series C*, vol. 36, pp. 318–324, 1987.

[182] S. Richardson and P. J. Green, "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 59, no. 4, pp. 731–792, 1997.

[183] Z. Zhang, C. Chen, J. Sun, and K. L. Chan, "EM algorithms for Gaussian mixtures with split-and-merge operation," *Pattern Recognition*, vol. 36, no. 9, pp. 1973–1983, 2003.

[184] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, dec 1974.

[185] R. J. Steele and E. Raftery Adrian, "Performance of Bayesian model selection criteria for Gaussian Mixture Model," Tech. Rep. 559, Department of statistics, University of Washington, 2009.

[186] A. D. Lanterman, "Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection," *International Statistical Review*, vol. 69, pp. 185–212, aug 2001.

[187] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for CBF activation studies in human brain," *Journal of Cerebral Blood Flow & Metabolism*, vol. 12, no. 900-918, 1992.

[188] K. J. Worsley, J.-B. B. Poline, a. C. Vandal, and K. J. Friston, "Tests for distributed, nonfocal brain activations," *NeuroImage*, vol. 2, pp. 183–195, 1995.

[189] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion.," *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, 2011.

[190] R. B. Fisher and P. Oliver, "Multi-variate cross-correlation and image matching," in *6th British Machine Vision Conference*, no. 1, pp. 623–632, 1995.

[191] C. H. Sudre, M. J. Cardoso, and S. Ourselin, "Bilayered anatomically constrained split-and-merge expectation maximisation algorithm (BiASM) for brain segmentation," in *SPIE Medical Imaging* (S. Ourselin and M. A. Styner, eds.), vol. 9034, pp. 903411–903411–7, International Society for Optics and Photonics, International Society for Optics and Photonics, 2014.

[192] G. Dugas-Phocion, M. Gonzalez, C. Lebrun, S. Chanalet, C. Bensa, G. Malandain, and N. Ayache, "Hierarchical segmentation of multiple sclerosis lesions in multi-sequence MRI," in *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (ISBI'04)*, pp. 157–160, 2004.

[193] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling, decomposition, and applications," *IEEE Transactions on Image Processing*, vol. 5, pp. 1293–1302, sep 1996.

[194] J. Hagemeier, J. J. G. Geurts, and R. Zivadinov, "Brain iron accumulation in aging and neurodegenerative disorders," *Expert Reviews Neurotherapeutics*, vol. 12, pp. 1467–1480, 2012.

[195] G. Gallego, C. Cuevos, R. Mohedano, and N. Garcia, "On the Mahalanobis distance classification criterion for multidimensional normal distribution," *IEEE Transactions on Signal Processing*, vol. 61, no. 17, pp. 4387–4396, 2013.

[196] C. H. Sudre, M. J. Cardoso, W. Bouvy, G. J. Biessels, J. Barnes, and S. Ourselin, "Bayesian Model Selection for Pathological Data," in *MICCAI 2014* (P. G. Et al., ed.), LNCS 8673, pp. 323–330, Springer International, 2014.

[197] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, and S. Ourselin, "Global image registration using a symmetric block-matching approach," *Journal of Medical Imaging*, vol. 1, no. 2, 2014.

[198] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin, "Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1976–1988, apr 2015.

[199] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin, "STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, no. 0, pp. 671–684, 2013.

[200] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[201] G. Borgefors, I. Nyström, G. di Baja, and G. Sanniti di Baja, "Connected components in 3D neighborhoods," in *10th Scandinavian Conference on Image Analysis*, pp. 567–572, jun 1997.

[202] D. S. Wack, M. G. Dwyer, N. Bergsland, C. Di Perri, L. Ranza, S. Hussein, D. Ramasamy, G. Poloni, and R. Zivadinov, "Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates," *BMC Medical imaging*, vol. 12, no. 17, 2012.

[203] J. de Bresser, A. M. Tiehuis, E. van den Berg, Y. D. Reijmer, C. Jongen, L. J. Kappelle, W. P. Mali, M. A. Viergever, and G. J. Biessels, "Progression of cerebral atrophy and white matter hyperintensities in patients with type 2 diabetes," *Diabetes Care*, vol. 33, pp. 1309–1314, jun 2010.

[204] C. Schwarz, E. Fletcher, C. DeCarli, and O. Carmichael, "Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR.," in *Information processing in medical imaging*, vol. 21, pp. 239–251, jan 2009.

[205] T. Tillin, N. G. Forouhi, P. M. McKeigue, N. f. t. S. group Chatuverdi, and N. Chaturvedi, "Southall And Brent REvisited: cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins," *International Journal of Epidemiology*, vol. 41, pp. 33–42, feb 2012.

[206] A. Ortiz, J. M. Górriz, J. Ramírez, and F. J. Martinez-Murcia, "Automatic ROI selection in structural brain MRI using SOM 3D projection," *PLoS ONE*, vol. 9, no. 4, 2014.

[207] K. K. Leung, J. Barnes, M. Modat, G. R. Ridgway, J. W. Bartlett, N. C. Fox, and S. Ourselin, "Brain MAPS: an automated, accurate and robust brain extraction technique using a template library," *NeuroImage*, vol. 55, pp. 1091–1108, apr 2011.

[208] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units.," *Computer Methods and Programs in Biomedicine*, vol. 98, pp. 278–284, jun 2010.

[209] D. F. Kallmes, F. K. Hui, and J. P. Mugler, "Suppression of cerebrospinal fluid and blood flow artifacts in FLAIR MR imaging with a single-slab three-dimensional pulse sequence: initial experience," *Radiology*, vol. 221, pp. 251–255, oct 2001.

[210] K. M. Hulsey, M. Gupta, K. S. King, R. M. Peshok, A. R. Whittermore, and R. W. McColl, "Automated quantification of white matter disease extent at 3T : Comparison with volumetric readings," *Journal of Magnetic Resonance Imaging*, vol. 36, pp. 305–311, 2012.

[211] M. Filippi, M. I. Gawne-Cain, C. Gasperini, J. H. van Waesberghe, J. Grimaud, F. Barkhof, M. P. Sormani, and D. H. Miller, "Effect of training and different measurement strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis," *Neurology*, vol. 50, pp. 238–244, 1998.

[212] S. Ropele, W. de Graaf, M. Khalil, M. P. Wattjes, C. Langkammer, M. A. Rocca, À. Rovira, J. Palace, F. Barkhof, M. Filippi, and F. Fazekas, "MRI asssessment of iron deposition in multiple sclerosis," *Journal of Magnetic Resonance Imaging*, vol. 34, pp. 13–21, 2011.

[213] M. de Groot, B. F. J. Verhaaren, R. de Boer, S. Klein, A. Hofman, A. van der Lugt, M. A. Ikram, W. J. Niessen, and M. W. Vernooij, "Changes in normal appearing white matter precede development of white matter lesions," *Stroke*, vol. 44, pp. 1037–1042, 2013.

[214] P. Maillard, O. Carmichael, E. Fletcher, B. Reed, D. Mungas, and C. DeCarli, "Coevolution of white matter hyperintensities and cognition in the elderly," *Neurology*, vol. 79, pp. 442–448, jul 2012.

[215] C. Enzinger, F. Fazekas, S. Ropele, and R. Schmidt, "Progression of cerebral white matter lesions : clinical and radiological considerations," *Journal of the Neurological Sciences*, vol. 257, pp. 5–10, jun 2007.

[216] A. M. Brickman, J. R. Sneed, F. a. Provenzano, E. Garcon, L. Johnert, J. Muraskin, L.-K. Yeung, M. E. Zimmerman, and S. P. Roose, "Quantitative approaches for assessment of white matter hyperintensities in elderly populations.," *Psychiatry Research*, vol. 193, pp. 101–106, 2011.

[217] F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman, "MR signal abnormalities on 1.5 T in Alzheimer's dementia and normal ageing," *American Journal of Neuroradiology*, vol. 8, pp. 421–426, 1987.

[218] P. Scheltens, F. Barkhof, D. Leys, J. P. Pruvo, J. J. P. Nauta, P. Vermersch, M. Steinling, J. Valk, V. P., M. Steinling, and J. Valk, "A semiquantitative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging," *Journal of the Neurological Sciences*, vol. 114, no. 1, pp. 7–12, 1993.

[219] F. Fazekas, R. Schmidt, and P. Scheltens, "Pathophysiologic Mechanisms in the Development of Age-Related White Matter Changes of the Brain," *Dementia and Geriatric Cognitive Disorders*, vol. 9, pp. 2–5, jul 1998.

[220] L. Bracco, C. Piccini, M. Moretti, M. Mascalchi, A. Sforza, B. Nacmias, E. Cellini, S. Bagnoli, and S. Sorbi, "Alzheimer's disease: role of size and location of white matter changes in determining cognitive deficits," *Dementia and Geriatric Cognitive Disorders*, vol. 20, pp. 358–366, jan 2005.

[221] F. Barkhof and P. Scheltens, "Is the whole brain periventricular?," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 77, pp. 143–144, feb 2006.

[222] F. van der Lijn, M. W. Vernooij, M. A. Ikram, H. A. Vrooman, D. Rueckert, A. Hammers, M. M. Breteler, and W. J. Niessen, "Automated localization of periventricular and subcortical white matter lesions," in *SPIE Medical Imaging*, pp. 651232–651232, International Society for Optics and Photonics, 2007.

[223] W. Wen and P. Sachdev, "The topography of white matter hyperintensities on brain MRI in healthy 60- to 64-year-old individuals.," *NeuroImage*, vol. 22, pp. 144–154, may 2004.

[224] F. van der Lijn, B. F. J. Verhaaren, M. A. Ikram, S. Klein, M. de Bruijne, H. A. Vrooman, M. W. Vernooij, A. Hammers, D. Rueckert, A. van der Lugt, M. M. B. Breteler, and W. J. Niessen, "Automated measurement of local white matter lesion volume," *NeuroImage*, vol. 59, pp. 3901–3908, 2012.

[225] S. E. Jones, B. R. Buchbinder, and A. Itzhak, "Three-dimensional mapping of cortical thickness using Laplace's equation," *Human Brain Mapping*, vol. 11, pp. 12–32, sep 2000.

[226] A. Yezzi and J. L. Prince, "A PDE approach for thickness, correspondence and gridding of annular tissues," in *EECV 2002* (A. Heyden, ed.), pp. 575–589, 2002.

[227] A. G. Jansen, S. E. Mous, T. White, D. Posthuma, and T. J. C. Polderman, "What twin studies tell us about the heritability of brain development, morphology, and function: a review," *Neuropsychology review*, vol. 25, pp. 27–46, mar 2015.

[228] J. S. Peper, R. M. Brouwer, D. I. Boomsma, R. S. Kahn, and H. E. Hulshoff Pol, "Genetic influences on human brain structure: a review of brain imaging studies in twins," *Human Brain Mapping*, vol. 28, pp. 464–473, jun 2007.

[229] R. M. Brouwer, R. C. W. Mandl, J. S. Peper, G. C. M. van Baal, R. S. Kahn, D. I. Boomsma, and H. E. Hulshoff Pol, "Heritability of DTI and MTR in nine-year-old children.," *NeuroImage*, vol. 53, pp. 1085–1092, nov 2010.

[230] A. Pfefferbaum, E. V. Sullivan, and D. Carmelli, "Morphological changes in aging brain structures are differentially affected by time-linked environmental influences despite strong genetic stability," *Neurobiology of Aging*, vol. 25, pp. 175–183, feb 2004.

[231] R. M. Brouwer, A. M. Hedman, N. E. M. van Haren, H. G. Schnack, R. G. H. Brans, D. J. A. Smit, R. S. Kahn, D. I. Boomsma, and H. E. Hulshoff Pol, "Heritability of brain volume change and its relation to intelligence.," *NeuroImage*, vol. 100, pp. 676–683, oct 2014.

[232] M. Dichgans, "Genetics of ischaemic stroke," *The Lancet Neurology*, vol. 6, pp. 149–161, feb 2007.

[233] D. Carmelli, C. DeCarli, G. E. Swan, L. M. Jack, T. Reed, P. A. Wolf, and B. L. Miller, "Evidence For Genetic Variance in White Matter Hyperintensity Volume in Normal Elderly Male Twins," *Stroke*, vol. 29, pp. 1177–1181, jun 1998.

[234] L. Harper, F. Barkhof, N. C. Fox, and J. M. Schott, "Using visual rating to diagnose dementia: a critical evaluation of MRI atrophy scales.," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 86, pp. 1225–1233, nov 2015.

[235] G. M. Potter, F. N. Doubal, C. A. Jackson, F. M. Chappell, C. L. Sudlow, M. S. Dennis, and J. M. Wardlaw, "Enlarged perivascular spaces and cerebral small vessel disease," *International Journal of Stroke*, vol. 10, pp. 376–81, apr 2015.

[236] H. J. Kuijf, S. J. van Veluw, M. A. Viergever, K. L. Vincken, and G. J. Biessels, "How to assess the reliability of cerebral microbleed rating?," *Frontiers in Aging Neuroscience*, vol. 5, jan 2013.

[237] W. T. Longstreth, T. A. Manolio, A. Arnold, G. L. Burke, N. Bryan, C. A. Jungreis, P. L. Enright, D. O'Leary, and L. Fried, "Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. The Cardiovascular Health Study," *Stroke*, vol. 27, pp. 1274–1282, aug 1996.

[238] C. Bocti, R. H. Swartz, F.-Q. Gao, D. J. Sahlas, P. Behl, and S. E. Black, "A new visual rating scale to assess strategic white matter hyperintensities within cholinergic pathways in dementia," *Stroke*, vol. 36, pp. 2126–2131, oct 2005.

[239] N. D. Prins, E. C. W. van Straaten, E. J. van Dijk, M. Simoni, R. a. van Schijndel, H. a. Vrooman, P. J. Koudstaal, P. Scheltens, M. M. B. Breteler, and F. Barkhof, "Measuring progression of cerebral white matter lesions on MRI: visual rating and volumetrics," *Neurology*, vol. 62, no. 9, pp. 1533–1539, 2004.

[240] T. A. Manolio, R. A. Kronmal, G. L. Burke, V. Poirier, D. H. O'Leary, J. M. Gardin, L. P. Fried, E. P. Steinberg, and R. N. Bryan, "Magnetic resonance abnormalities and cardiovascular disease in older adults - The cardiovascular Health study," *Stroke*, vol. 25, no. 2, pp. 318–327, 1994.

[241] C. Oboudiyat, H. Gardener, C. Marquez, M. Elkind, R. Sacco, C. DeCarli, and C. Wright, "Comparing Semi-quantitative and Volumetric Measurements of MRI White Matter Hyperintensities: The Northern Manhattan Study (S62.007)," *Neurology*, vol. 82, pp. S62.007–, apr 2014.

[242] J. M. Stankiewicz, B. I. Glanz, B. C. Healy, A. Arora, M. Neema, R. H. B. Benedict, Z. D. Guss, S. Tauhid, G. J. Buckle, M. K. Houtchens, S. Khoury, H. L. Weiner, C. R. G. Guttmann, and R. Bakshi, "Brain MRI lesion load at 1.5T and 3T vs. clinical status in multiple sclerosis," *Journal of Neuroimaging*, vol. 21, no. 2, pp. 1–15, 2011.

[243] R. Schmidt, S. Seiler, and M. Loitfelder, "Longitudinal change of small-vessel disease related brain abnormalities," *Journal of Cerebral Blood Flow & Metabolism*, no. February, pp. 1–8, 2015.

[244] E. J. van Dijk, N. D. Prins, H. A. Vrooman, A. Hofman, P. J. Koudstaal, and M. M. B. Breteler, "Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam Scan study," *Stroke*, vol. 39, pp. 2712–2719, oct 2008.

[245] L. C. Silbert, C. Nelson, D. B. Howieson, M. M. Moore, and J. A. Kaye, "Impact of white matter hyperintensity volume progression on rate of cognitive and motor decline," *Neurology*, vol. 71, pp. 108–113, jul 2008.

[246] Z. Liu, Y. Zhao, H. Zhang, Q. Chai, Y. Cui, Y. Diao, J. Xiu, X. Sun, and G. Jiang, "Excessive variability in systolic blood pressure that is self-measured at home exacerbates the progression of brain white matter lesions and cognitive impairment in the oldest old," *Hypertension Research*, dec 2015.

[247] D. Inzitari, G. Pracucci, A. Poggesi, G. Carlucci, F. Barkhof, H. Chabriat, T. Erkinjuntti, F. Fazekas, J. M. Ferro, M. Hennerici, P. Langhorne, J. O'Brien, P. Scheltens, M. C. Visser, L.-O. Wahlund, G. Waldemar, A. Wallin, and L. Pantoni, "Changes in white matter as determinant of global functional decline in older independent outpatients: three year follow-up of LADIS (leukoaraiosis and disability) study cohort," *BMJ*, vol. 339, p. b2477, jan 2009.

[248] A. a. Gouw, W. M. Van Der Flier, E. C. W. van Straaten, L. Pantoni, A. J. Bastos-Leite, D. Inzitari, T. Erkinjuntti, L.-O. O. Wahlund, C. Ryberg, R. Schmidt, F. Fazekas, P. Scheltens, and F. Barkhof, "Reliability and sensitivity of visual scales versus volumetry for evaluating white matter hyperintensity progression," *Cerebrovascular Diseases*, vol. 25, pp. 247–253, 2008.

[249] R. Schmidt, K. Petrovic, S. Ropele, C. Enzinger, and F. Fazekas, "Progression of leukoaraiosis and cognition," *Stroke*, vol. 38, pp. 2619–2625, sep 2007.

[250] M. d. C. Valdés Hernández, V. González-Castro, D. T. Ghandour, X. Wang, F. Doubal, S. Muñoz Maniega, P. A. Armitage, and J. M. Wardlaw, "On the computational assessment of white matter hyperintensity progression: difficulties in method selection and bias field correction performance on images with significant white matter pathology," *Neuroradiology*, jan 2016.

[251] N. C. Fox, G. R. Ridgway, and J. M. Schott, "Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials," *NeuroImage*, vol. 57, pp. 15–8, jul 2011.

[252] M. Reuter and B. Fischl, "Avoiding asymmetry-induced bias in longitudinal image processing," *NeuroImage*, vol. 57, pp. 19–21, jul 2011.

[253] C. Elliott, D. L. Arnold, D. L. Collins, and T. Arbel, "Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1490–1503, 2013.

[254] Y. Duan, P. Hildenbrand, M. Sampat, D. Tate, I. Csapo, B. Moraal, R. Bakshi, F. Barkhof, D. Meier, and C. Guttmann, "Segmentation of Subtraction Images for the Measurement of Lesion Change in Multiple Sclerosis," *American Journal of Neuroradiology*, vol. 29, pp. 340–346, feb 2008.

[255] O. Ganiler, A. Oliver, Y. Diez, J. Freixenet, J. C. Vilanova, B. Beltran, L. Ramió-Torrentà, À. Rovira, X. Lladó, A. Rovira, and X. Lladó, "A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies," *Neuroradiology*, vol. 56, pp. 363–374, may 2014.

[256] D. Rey, G. Subsol, H. Delingette, and N. Ayache, "Automatic detection and segmentation of evolving processes in 3D medical images: Application to multiple sclerosis," *Medical Image Analysis*, vol. 6, pp. 163–179, jun 2002.

[257] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach, "Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution," *NeuroImage*, vol. 20, pp. 643–656, 2003.

[258] C. Elliott, D. L. Arnold, D. L. Collins, and T. Arbel, "A generative model for automatic detection of resolving multiple sclerosis lesions," in *First International Workshop on Bayesian and grAphical Models for Biomedical Imaging*, pp. 118–129, 2014.

[259] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[260] M. Prastawa, E. Bullitt, and G. Gerig, "Simulation of brain tumors in MR images for evaluation of segmentation efficacy," *Medical image analysis*, vol. 13, pp. 297–311, apr 2009.

[261] E. R. Melhem, E. H. Herskovits, K. Karli-Oguz, X. Golay, D. A. Hammoud, B. J. Fortman, F. M. Munter, and R. Itoh, "Defining thresholds for changes in size of simulated t2-hyperintense brain lesions on the basis of qualitative comparisons," *American Journal of Roentgenology*, vol. 180, no. 1, pp. 65–69, 2003.

[262] C. Sudre, M. J. Cardoso, W. Bouvy, G. Biessels, J. Barnes, and S. Ourselin, "Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation.," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 2079–2102, apr 2015.

[263] B. V. Zlokovic, "Cerebrovascular Effects of Apolipoprotein E -Implications for Alzheimer Disease," *JAMA Neurology*, vol. 70, p. 1, apr 2013.

[264] P. B. Verghese, J. M. Castellano, and D. M. Holtzman, "Apolipoprotein E in Alzheimer's disease and other neurological disorders," *The Lancet Neurology*, vol. 10, pp. 241–252, mar 2011.

[265] J. E. Donahue and C. E. Johanson, "Apolipoprotein E, amyloid-beta, and blood-brain barrier permeability in Alzheimer disease," *Journal of Neuropathology & Experimental Neurology*, vol. 67, no. 4, pp. 261–270, 2008.

[266] J. Kim, J. M. Basak, and D. M. Holtzman, "The role of apolipoprotein E in Alzheimer's disease.," *Neuron*, vol. 63, pp. 287–303, aug 2009.

[267] A. Gupta and C. Iadecola, "Impaired ABeta clearance: a potential link between atherosclerosis and Alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 7, 2015.

[268] B. V. Zlokovic, "Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders.," *Nature reviews. Neuroscience*, vol. 12, pp. 723–38, dec 2011.

[269] L. M. Tai, R. Thomas, F. M. Marottoli, K. P. Koster, T. Kanekiyo, A. W. J. Morris, and G. Bu, "The role of APOE in cerebrovascular dysfunction.," *Acta Neuropathologica*, feb 2016.

[270] K. Nishitsuji, T. Hosono, T. Nakamura, G. Bu, and M. Michikawa, "Apolipoprotein E regulates the integrity of tight junctions in an isoform-dependent manner in an in vitro blood-brain barrier model.," *Journal of Biological Chemistry*, vol. 286, pp. 17536–17542, may 2011.

[271] R. D. Bell, E. A. Winkler, I. Singh, A. P. Sagare, R. Deane, Z. Wu, D. M. Holtzman, C. Betsholtz, A. Armulik, J. Sallstrom, B. C. Berk, and B. V. Zlokovic, "Apolipoprotein E controls cerebrovascular integrity via cyclophilin A," *Nature*, vol. 485, pp. 512–516, may 2012.

[272] W. Alata, Y. Ye, I. St-Amour, M. Vandal, and F. Calon, "Human apolipoprotein E 4 expression impairs cerebral vascularization and blood–brain barrier function in mice," *Journal of Cerebral Blood Flow & Metabolism*, vol. 35, pp. 86–94, jan 2014.

[273] S. Salloway, T. Gur, T. Berzin, B. Zipser, S. Correia, V. Hovanesian, J. Fallon, V. Kuo-Leblanc, D. Glass, C. Hulette, C. Rosenberg, M. Vitek, and E. Stopa, "Effect of APOE genotype on microvascular basement membrane in Alzheimer's disease," *Journal of the Neurological Sciences*, vol. 203-204, pp. 183–187, nov 2002.

[274] P. A. Yates, V. L. Villemagne, K. A. Ellis, P. M. Desmond, C. L. Masters, and C. C. Rowe, "Cerebral microbleeds: a review of clinical, genetic, and neuroimaging associations.," *Frontiers in Neurology*, vol. 4, p. 205, jan 2014.

[275] Y. Song, "Meta-Analysis: Apolipoprotein E Genotypes and Risk for Coronary Heart Disease," *Annals of Internal Medicine*, vol. 141, p. 137, jul 2004.

[276] T. A. Khan, T. Shah, D. Prieto, W. Zhang, J. Price, G. R. Fowkes, J. Cooper, P. J. Talmud, S. E. Humphries, J. Sundstrom, J. A. Hubacek, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, M. R. Abdollahi, A. J. C. Slooter, Z. Szolnoki, M. Sandhu, N. Wareham, R. Frikke-Schmidt, A. Tybjærg-Hansen, G. Fillenbaum, B. T. Heijmans, T. Katsuya, G. Gromadzka, A. Singleton, L. Ferrucci, J. Hardy, B. Worrall, S. S. Rich, M. Matarin, J. Whittaker, T. R. Gaunt, P. Whincup, R. Morris, J. Deanfield, A. Donald, G. Davey Smith, M. Kivimaki, M. Kumari, L. Smeeth, K.-T. Khaw, M. Nalls, J. Meschia, K. Sun, R. Hui, I. Day, A. D. Hingorani, and J. P. Casas, "Apolipoprotein E genotype, cardiovascular biomarkers and risk of stroke: systematic review and meta-analysis of 14,015 stroke cases and pooled analysis of primary biomarker data from up to 60,883 individuals.," *International Journal of Epidemiology*, vol. 42, pp. 475–92, apr 2013.

[277] F. E. De Leeuw, F. Richard, J. C. De Groot, C. M. Van Duijn, A. Hofman, J. Van Gijn, and M. M. B. Breteler, "Interaction between hypertension, apoE, and cerebral white matter lesions," *Stroke*, vol. 35, no. 5, pp. 1057–1060, 2004.

[278] R. Wang, L. Fratiglioni, E. J. Laukka, M. Lövdén, L. Keller, and C. Graff, "Effects of vascular risk factors and APOE e 4 on white matter integrity and cognitive decline," *Neurology*, vol. 84, pp. 1128–1135, 2015.

[279] S. Schilling, A. L. DeStefano, P. S. Sachdev, S. H. Choi, K. A. Mather, C. D. DeCarli, W. Wen, P. Høgh, N. Raz, R. Au, A. Beiser, P. A. Wolf, J. R. Romero, Y.-C. Zhu, K. L. Lunetta, L. Farrer, C. Dufouil, L. H. Kuller, B. Mazoyer, S. Seshadri, C. Tzourio, and S. Debette, "APOE genotype and MRI markers of cerebrovascular disease: systematic review and meta-analysis," *Neurology*, vol. 81, pp. 292–300, jul 2013.

[280] H. J. Kim, B. S. Ye, C. W. Yoon, H. Cho, Y. Noh, G. H. Kim, Y. S. Choi, J. H. Kim, S. Jeon, J. M. Lee, J. S. Kim, Y. S. Choe, K. H. Lee, S. T. Kim, C. Kim, D. R. Kang, C. S. Ki, J. H. Lee, D. J. Werring, M. W. Weiner, D. L. Na, and S. W. Seo, "Effects of APOE ε4 on brain amyloid, lacunar infarcts, and white matter lesions: Astudy among patients with subcortical vascular cognitive impairment," *Neurobiology of Aging*, vol. 34, no. 11, pp. 2482–2487, 2013.

[281] Y. Hoi, Y. Q. Zhou, X. Zhang, R. M. Henkelman, and D. A. Steinman, "Correlation between local hemodynamics and lesion distribution in a novel aortic regurgitation murine model of atherosclerosis," *Annals of Biomedical Engineering*, vol. 39, no. 5, pp. 1414–1422, 2011.

[282] V. Heise, N. Filippini, K. P. Ebmeier, and C. E. Mackay, "The APOE ε4 allele modulates brain white matter integrity in healthy adults," *Molecular Psychiatry*, vol. 16, no. 9, pp. 908–916, 2011.

[283] L. Nyberg and A. Salami, "The APOE 4 allele in relation to brain white-matter microstructure in adulthood and aging," *Scandinavian Journal of Psychology*, vol. 55, no. 3, pp. 263–267, 2014.

[284] J. Ma, A. Yee, H. B. Brewer, S. Das, and H. Potter, "Amyloid-associated proteins alpha 1-antichymotrypsin and apolipoprotein E promote assembly of Alzheimer beta-protein into filaments.," *Nature*, vol. 372, pp. 92–94, nov 1994.

[285] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media, 2000.

[286] P. Sachdev, W. Wen, X. Chen, and H. Brodaty, "Progression of white matter hyperintensities in elderly individuals over 3 years," *Neurology*, vol. 68, pp. 214–222, jan 2007.

[287] O. Godin, C. Tzourio, P. Maillard, A. Alpérovitch, B. Mazoyer, and C. Dufouil, "Apolipoprotein E genotype is related to progression of white matter lesion load," *Stroke*, vol. 40, pp. 3186–90, oct 2009.

[288] R. Y. Lo and W. J. Jagust, "Vascular burden and Alzheimer disease pathologic progression," *Neurology*, vol. 79, pp. 1349–1355, sep 2012.

[289] N. Andreasen, C. Hesse, P. Davidsson, L. Minthon, A. Wallin, B. Winblad, H. Vanderstichele, E. Vanmechelen, and K. Blennow, "Cerebrospinal Fluid β-Amyloid(1-42) in Alzheimer Disease," *Archives of Neurology*, vol. 56, p. 673, jun 1999.

[290] V. Stenset, L. Johnsen, D. Kocot, A. Negaard, A. Skinningsrud, P. Gulbrandsen, A. Wallin, and T. Fladby, "Associations between white matter lesions, cerebrovascular risk factors, and low CSF Abeta42," *Neurology*, vol. 67, pp. 830–833, sep 2006.

[291] B. Gesierich, C. Opherk, J. Rosand, M. Gonik, R. Malik, E. Jouvent, D. Hervé, P. Adib-Samii, S. Bevan, L. Pianese, S. Silvestri, M. T. Dotti, N. De Stefano, J. van der Grond, E. M. Boon, F. Pescini, N. Rost, L. Pantoni, S. a. Lesnik Oberstein, A. Federico, M. Ragno, H. S. Markus, E. Tournier-Lasserve, H. Chabriat, M. Dichgans, M. Duering, and M. Ewers, "APOE 2 is associated with white matter hyperintensity volume in CADASIL," *Journal of Cerebral Blood Flow & Metabolism*, no. April, pp. 1–4, 2015.

[292] B. H. Menze, K. Van Leemput, D. Lashkari, T. Riklin-Raviv, E. Geremia, E. Alberts, P. Gruber, S. Wegener, M.-A. Weber, G. Székely, *et al.*, "A generative probabilistic model and discriminative extensions for brain lesion segmentation—with application to tumor and stroke," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 933–946, 2016.